Inferences for Two-Stage Multiple Imputation for Nonresponse

S. K. Kinney, *National Institute of Statistical Sciences,*
Research Triangle Park, NC 27709, USA
Email: saki@niss.org
J. P. Reiter, *Department of Statistical Science,*
Duke University, Durham, NC 27708, USA
Email: jerry@stat.duke.edu

Abstract

Multiple imputation is a common approach for handling missing data. It allows users to make valid inferences using standard complete-data methods with simple combining rules. A variation is to partition the missing data into two portions and conduct the imputation in two stages. We review two-stage multiple imputation and existing inferential methods and derive an alternative reference F-distribution for large sample hypothesis testing for high-dimensional estimands. We also derive formulas for estimating rates of missing information.

Key Words: Multiple imputation, Nonresponse, Significance testing, Surveys.

1. Introduction

Multiple imputation was first proposed for handling nonresponse in large complex surveys. The goal was to facilitate valid inferences when the data producer and the ultimately many end users of the data were distinct entities. In this scenario, the burden of modeling the missing data mechanism lies on the data producer, who may have skills and information unavailable to the users, while the users are able to focus on their analyses without learning new or complex missing data methods (Rubin, 1996). Multiple imputation is now commonly used to handle missing data by agencies as well as individual users. Several software packages, including R, SAS, and SPlus, have routines that simplify the process for both filling in missing values with multiple imputations and drawing inferences from completed datasets. In addition to missing data, multiple imputation is now used in other applications, including statistical disclosure limitation (Rubin, 1993; Little, 1993; Reiter, 2005, 2003) and measurement error (Clogg *et al.*, 1991; Cole *et al.*, 2006). These are reviewed in Reiter and Raghunathan (2007).

Nested or two-stage imputation refers to multiple imputation conducted in a nested fashion. In the first stage, m imputations are generated. In the second stage, n imputations are generated for each completed data set in the first stage, resulting in a total of $M = mn$ multiply-imputed data sets. Nested multiple imputation was first proposed in Shen (2000), motivated in part by the multiple imputation of missing data in the National Medical Expenditure Survey. In this project, a large number of imputations were generated, with reduced computational burden, by splitting the missing data into two parts, where one part was computationally intensive and the other computationally inexpensive. First, a small number of imputations were generated for the computationally intensive portion, which included all the data except medical expenditures with missing disease codes. These took ten days per imputation to create. Then, for each imputed dataset, several imputations were generated for the inexpensive portion comprising the missing disease codes and the associated expenditures. Releasing M imputations reduced variances relative to releasing only m imputations (Rubin, 2003; Shen, 2000).

In addition to reducing computational burdens, two-stage imputation is useful when imputation of one partition would be substantially easier if the other were known and when different numbers of imputations are desired for two partitions of missing data. Additionally, it is often the case that missing data are of different types, such as planned and unplanned nonresponse, which contribute qualitatively different types of variability. While one-stage imputation may still be used in these cases, the use of two-stage imputation can result in inferences with reduced variances. Two-stage imputation also can enable imputers to isolate the effects of different types of missingness, evaluate different sources of variability, and measure the expected increase in information if one part were known. These can inform future studies (Harel and Schafer, 2003).

Two-stage imputation has been applied in applications of multiple imputation other than handling missing data. Reiter and Drechsler (2007) show that two-stage imputation can reduce computational burdens when using multiple imputation for statistical disclosure control. They also use two-stage imputation to release fewer imputations for variables at high risk of disclosure than for variables at low risk of disclosure. Reiter (2007) uses two-stage imputation to enable valid inferences when some of the records used to generate the imputations are not made available to the analyst. An example of this is when multiple imputation is applied to address measurement error using external validation data that are not released. Reiter (2004) uses two-stage imputation approach to address disclosure limitation and missing data simultaneously. Additional uses of two-stage multiple imputation are suggested in Harel and Schafer (2003) and Reiter and Raghunathan (2007).

The remainder of this article focuses on two-stage multiple imputation for missing data. Section 2 reviews two-stage multiple imputation and existing inferential methods. Section 3 presents an improved multivariate test for high-dimensional estimands. Section 4 illustrates the improved performance of the proposed test with simulations. Section 5 reviews the work of Harel and Schafer (2003) on rates of missing information and extends to multivariate estimands with finite M.

2. Review and notation

For a finite population of size N, let $I_l = 1$ if unit l is selected in the survey, and $I_l = 0$ otherwise, where $l = 1, \ldots, N$. Let $I = (I_1, \ldots, I_N)$, and let the sample size $s = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables, such as stratum or cluster indicators, and assume that X is known at least approximately for the population. Let Y be the $N \times p$ matrix of survey data for the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $s \times p$ sub-matrix of Y for units with $I_l = 1$, where Y_{obs} is the portion that is observed and Y_{mis} is the portion that is missing due to nonresponse. Let $R = (R^{(A)}, R^{(B)})$, where $R^{(A)}$ is an $N \times p$ matrix of indicators such that $R^{(A)}_{lk} = 1$ if the response for unit l to item k is missing and to be imputed in the first stage and $R^{(A)}_{lk} = 0$ otherwise, and $R^{(B)}$ be the corresponding $N \times p$ matrix of indicators for the second stage of imputation and partition Y_{mis} into $Y^{(A)}_{mis}$ and $Y^{(B)}_{mis}$.

To generate imputations, the imputer first fills in $Y^{(A)}_{mis}$ with m draws from the posterior distribution of $(Y^{(A)}_{mis}|D_{obs})$, resulting in m partially completed datasets, $D_{pcom} = \{D^{(i)}_{pcom}, i = 1, \ldots, m\}$, where $D^{(i)}_{pcom}$ is comprised of D_{obs}, $Y^{(B)}_{mis}$, and the ith imputation of $Y^{(A)}_{mis}$. Then, for each $D^{(i)}_{pcom}$, the imputer fills in $Y^{(B)}_{mis}$ with n draws from the posterior predictive distribution of $(Y^{(B)}_{mis}|D^{(i)}_{pcom}, R^{(B)})$, resulting in a total of $M = mn$ imputed datasets $D_{com} = \{D^{(i,j)}_{com}, i = 1, \ldots, m; j = 1, \ldots, n\}$, where $D^{(i,j)}_{com}$ is comprised of D_{obs}, the ith imputation of $Y^{(A)}_{mis}$ and the jth imputation of $Y^{(B)}_{mis}$.

Shen (2000) describes a second, equivalent method of generating imputations in two stages. In this procedure, m imputations of $Y^{(A)}_{mis}$ and $Y^{(B)}_{mis}$ are drawn in the first stage from the joint posterior distribution of $(Y^{(A)}_{mis}, Y^{(B)}_{mis}|D_{obs})$. In the second stage, an additional $n - 1$ conditionally independent imputations are drawn from the distribution of $(Y^{(B)}_{mis}|D^{(i)}_{pcom}, R^{(B)})$. This approach is advantageous when it is easier to specify or draw from the distribution of $(Y^{(A)}_{mis}, Y^{(B)}_{mis}|D_{obs})$ than from the distribution of $(Y^{(A)}_{mis}|D_{obs})$.

When nested imputation is used for the purpose of reducing computational efforts, the computationally intensive portion is naturally chosen to be imputed first so as to minimize the number of imputations, so that $m < n$. Absent computational concerns, for randomization validity it makes sense to impute the portion with a greater proportion of missing values first, so that $m > n$. Harel (2003) suggests setting $n = 2$ and choosing m to obtain the desired precision, unless the rate of missing information in the first stage is thought to be much smaller than in the second. In the similar setting of two-stage imputation for missing data and disclosure limitation, Reiter (2008) found improved inferences when $m > n$, particularly for large fractions of missing data in the first stage.

Estimates based on two-stage multiple imputation can have smaller or larger variances than those based on one-stage imputation. It depends on how one makes the comparison. Compared to one-stage imputation with m data sets, two-stage imputation with M data sets provides more information (lower variances) for estimates that depend on $Y_{mis}^{(B)}$; there are no differences in the variances of estimates that depend only on $Y_{mis}^{(A)}$. Compared to one stage imputation with M data sets, two stage imputation provides less information (higher variances). This is because the second stage imputations are correlated and represent fewer than M independent pieces of information.

When the imputer believes that r one-stage imputations provide adequate precision, but for computational reasons wants to generate $m < r$ imputations in the first stage of two-stage imputation, the imputer must set $M > r$ to achieve similar precision (for estimates that depend on $Y_{mis}^{(B)}$). Results from Shen (2000) suggest the reduction in computational effort for the imputations in the first stage is offset somewhat by the need for more imputations in the second stage to achieve a precision similar to r imputations in one-stage imputation.

2.1 Existing inferential methods

Shen (2000) develops a combining rule for univariate estimands and derives a test for multicomponent estimands, noting that the analytic validity does not hold when the dimension of the estimand is high relative to the number of imputations. Valid inferences for multiply-imputed data are obtained for a parameter Q by obtaining standard complete-data estimates from each completed dataset $D_{com}^{(i,j)}$ and applying simple combining rules. The combining rules for one-stage multiple imputation of Rubin (1987) do not apply to data imputed in two stages, as the imputations are not exchangeable. The following quantities are needed for inferences about some k-dimensional parameter Q:

$$\bar{Q} = \sum_{i=1}^{m}\sum_{j=1}^{n} Q^{(i,j)}/mn = \sum_{i=1}^{m} \bar{Q}^{(i)}/m \tag{2.1}$$

$$\bar{U} = \sum_{i=1}^{m}\sum_{j=1}^{n} U^{(i,j)}/mn \tag{2.2}$$

$$\bar{W} = \sum_{i=1}^{m}\sum_{j=1}^{n} (Q^{(i,j)} - \bar{Q}^{(i)})^2/m(n-1) = \sum_{i=1}^{m} W^{(i)}/m \tag{2.3}$$

$$B = \sum_{i=1}^{m} (\bar{Q}^{(i)} - \bar{Q})^2/(m-1) \tag{2.4}$$

where $\bar{Q}^{(i)}$ is the average of the point estimates in the nest of datasets indexed by i, $\bar{Q}$ is the average of the $\bar{Q}^{(i)}$ across nests; $W^{(i)}$ is the within-group variances of the point estimates in the nest of datasets indexed by i, and $\bar{W}$ is the average of the $W^{(i)}$; B is the between-group variance of the $\bar{Q}^{(i)}$ across nests; and, $\bar{U}$ is the average of the estimated variances of $Q^{(i,j)}$ across all imputed datasets.

Inferences for some scalar parameter q are based on the quantities in (2.1) to (2.4) with $k = 1$. The estimate of q is $\bar{q}$, and the variance of $\bar{q}$ is $T_n = (1 + 1/m)b + (1 - 1/n)\bar{w} + \bar{u}$, where lower-case letters

denote scalar quantities. If $\bar{w} = 0$ or $n = 1$, T_n reduces to $T_m = \bar{u} + (1 + 1/m)b$, the standard variance estimate for one-stage multiple imputation (Rubin, 1987). When the sample size s is sufficiently large, inferences for q can be based on t-distributions with mean $\bar{q}$, variance T_n and degrees of freedom $\nu_n = \left\{ \frac{((1+1/m)b)^2}{(m-1)T_n^2} + \frac{((1-1/n)\bar{w})^2}{m(n-1)T_n^2} \right\}^{-1}$.

With one stage multiple imputation, T_m can be biased in some settings (Wang and Robins, 1998; Robins and Wang, 2000; Nielsen, 2003; Kim *et al.*, 2006). However, Rubin (2003) and others argue that the bias typically is not substantial enough to outweigh the benefits of using T_m and multiple imputation in general. Similar results hold for T_n in two stage imputation. For example, Shen (2000) demonstrates that T_n results in inferences with good frequentist properties in a variety of settings.

2.2 Inferences for multivariate parameters

Generalizing from the univariate case, let Q be a multicomponent estimand, such as a vector of regression coefficients. The quantities in (2.1) through (2.4) are used for inferences about Q, with the expected value given by $\bar{Q}$. An estimate of the variance of $\bar{Q}$ is given by $T_n = (1 + 1/m)B + (1 - 1/n)W + \bar{U}$.

When testing $H_0 : Q = Q_0$, for multivariate parameter Q, it may seem reasonable to use a Wald test with test statistic $(Q_0 - \bar{Q})T_n^{-1}(Q_0 - \bar{Q})$ when the sample size s is sufficiently large; however, T_n can be a poor estimate of the variance when m and n are modest. Estimating B and $\bar{W}$ for modest values of m and n is akin to estimating the covariance matrix with few observations relative to the dimension. Hence, tests based on this covariance estimate perform poorly in cases of practical interest, and a modification is needed.

When the covariance matrices $U^{(i,j)}$ are available, we use the test statistic

$$S_n = (Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})/k(1 + r_n^{(b)} + r_n^{(w)}) \tag{2.5}$$

where

$$r_n^{(b)} = (1 + 1/m)\text{tr}(B\bar{U}^{-1})/k \tag{2.6}$$
$$r_n^{(w)} = (1 - 1/n)\text{tr}(\bar{W}\bar{U}^{-1})/k. \tag{2.7}$$

Shen (2000) proposes an approximate Bayesian p-value extending the approach of Rubin (1987). This is obtained by referring S_n to an F_{k,w_n^*} distribution, where

$$w_n^* = \left\{ \frac{(r_n^{(b)})^2}{\nu_b(1 + r_n^{(b)} + r_n^{(w)})^2} + \frac{(r_n^{(w)})^2}{\nu_w(1 + r_n^{(b)} + r_n^{(w)})^2} \right\}^{-1} \tag{2.8}$$

and $\nu_b = k(m - 1)$ and $\nu_w = km(n - 1)$.

Some software programs may not make covariance matrices of parameter estimates readily available, and $\bar{U}$ may be unwieldy for large k. Meng and Rubin (1992) developed an alternative test for conventional multiply-imputed data for missing data, based on the set of log-likelihood ratio test statistics from a set of completed datasets. These do not require any U_{ij} and are easily computed for common models appropriate for the standard combining rules. Shen (2000) also extended this approach to two-stage imputation.

The basis of the approach is the asymptotic equivalence of the Wald and log likelihood ratio test statistics. A test statistic $\tilde{S}_n$ is found that is asymptotically equivalent to S_n and can be computed with access only to the Wald statistics calculated using each individual synthetic dataset. The asymptotic relationship between the Wald and log likelihood ratio statistics is used to obtain the test statistic, $\tilde{S}_n$, and denominator degrees of freedom $\tilde{w}_n^*$. The test is conducted by referring $\tilde{S}_n$ to an $F_{k,\tilde{w}_n^*}$-distribution. For details of the test and its derivation, see Shen (2000).

3. Proposed test for multivariate estimands

Shen (2000) found that the test based on S_n and w_n^* exhibited poor frequentist properties when k was large relative to m. The corresponding test for single-stage multiple imputation is known to have the same problem. Li *et al.* (1991a) proposed an alternate denominator degrees of freedom to that of Rubin (1987) for one-stage multiple imputation that has better frequentist properties. It is widely used for testing multicomponent hypotheses. We extend this approach to two-stage multiple imputation and use a new denominator degrees of freedom given by:

$$w_n = 4 + \left\{1 + \frac{r_n^{(b)}\nu_b}{\nu_b - 2} + \frac{r_n^{(w)}\nu_w}{\nu_w - 2}\right\}^2 / \left\{\frac{(r_n^{(b)}\nu_b)^2}{(\nu_b-2)^2(\nu_b-4)} + \frac{(r_n^{(w)}\nu_w)^2}{(\nu_w-2)^2(\nu_w-4)}\right\}. \tag{3.1}$$

When $\bar{W} = 0$ or $n = 1$, S_n and w_n reduce to the test statistic and degrees of freedom for missing data imputed in one stage (Li *et al.*, 1991a). Similarly, the likelihood ratio test of Shen (2000), described in Section 2.2, is based on w_n rather than w_n^*.

When $\nu_b \leq 4$ or $\nu_w \leq 4$, w_n is not defined; however, this only occurs for cases with $m = 2$ and k small. When a user is faced with a situation where w_n is undefined, the test is based on w_n^*.

3.1 Derivation

The derivation given here for the test statistic S_n is similar to that presented in Shen (2000); however, the derivation of the reference distribution is substantially different. Most notably, we do not ignore the lack of independence between the variance parameters corresponding to the between-nest and within-nest variances.

Let $B_\infty = \lim B$ as $m \to \infty$ and $n \to \infty$; let $\bar{W}_\infty = \sum W_\infty^{(i)}/m$ where $W_\infty^{(i)} = \lim W^{(i)}$ as $n \to \infty$; and, let $\bar{U}_\infty = \lim \bar{U}$ as $m \to \infty$ and $n \to \infty$. Assuming the conditions for valid inferences under multiple imputation (Rubin, 1987; Harel, 2003), the posterior distribution of $(Q|D_{com}, B_\infty, \bar{W}_\infty)$ is $N(\bar{Q}, T_\infty)$, where $T_\infty = \bar{U}_\infty + (1+1/m)B_\infty + (1+1/mn)\bar{W}_\infty$. If T_∞ were known, then the Bayesian p-value for testing $H_0 : Q = Q_0$ would be $P(\chi_k^2 > (Q_0 - \bar{Q})'T_\infty^{-1}(Q_0 - \bar{Q}))$. Since T_∞ is generally not known, the p-value is obtained by integrating over the conditional distributions of the variance parameters $(B_\infty|D_{com}, \bar{W}_\infty)$ and $(\bar{W}_\infty|D_{com})$:

$$\iint P\{\chi_k^2 > (Q_0 - \bar{Q})'T_\infty^{-1}(Q_0 - \bar{Q})|D_{com}, B_\infty, \bar{W}_\infty\} \times P(B_\infty|D_{com}, \bar{W}_\infty)P(\bar{W}_\infty|D_{com})dB_\infty d\bar{W}_\infty. \tag{3.2}$$

To obtain a closed-form approximation, and to reduce the number of variance parameters to be estimated, we assume that the between-nest variance B_∞ and within-nest variance $\bar{W}_\infty$ are both proportional to the total variance and hence to $\bar{U}_\infty$:

$$B_\infty = r_\infty^{(b)}\bar{U}_\infty, \bar{W}_\infty = r_\infty^{(w)}\bar{U}_\infty \tag{3.3}$$

for scalar quantities $r_\infty^{(w)}$ and $r_\infty^{(b)}$, not assumed to be equal. That is, for each stage, we assume equal fractions of missing information (which can differ by stage) for each component of Q. Under (3.3), (3.2) reduces to

$$\iint P\left\{\chi_k^2 > \frac{(Q_0 - \bar{Q})'U_\infty^{-1}(Q_0 - \bar{Q})}{1 + (1 + \frac{1}{m})r_\infty^{(b)} + (1 + \frac{1}{mn})r_\infty^{(w)}}|D_{com}\right\} \times P(r_\infty^{(b)}|D_{com}, r_\infty^{(w)})P(r_\infty^{(w)}|D_{com})dr_\infty^{(b)}dr_\infty^{(w)}. \tag{3.4}$$

Under asymptotic theory for the sampling distribution of the posterior variance, which tends to have lower posterior variance than the mean, $\bar{U}_\infty$ can be replaced with $\bar{U}$ (Rubin, 1987, p.89). Generalizing from the

theory for univariate estimands, we have

$$\begin{aligned}\{B|(B_\infty + \bar{W}_\infty/n)^{-1}|D_{com}, \bar{W}_\infty)\} &\sim Wish(m-1, I)\\ \{\bar{W}_\infty(\bar{W}_\infty)^{-1}|D_{com})\} &\sim Wish(m(n-1), I).\end{aligned}$$

Assuming (3.3), applying standard multivariate normal theory, and averaging across nests, the conditional distributions of $r_\infty^{(b)}$ and $r_\infty^{(w)}$ follow as:

$$\begin{aligned}\left\{\frac{k(m-1)\text{tr}(B\bar{U}^{-1})/k}{r_\infty^{(b)} + r_\infty^{(w)}/n}|D_{syn}, r_\infty^{(w)}\right\} &\sim \chi^2_{k(m-1)}\\ \left\{\frac{km(n-1)\text{tr}(\bar{W}\bar{U}^{-1})/k}{r_\infty^{(w)}}|D_{syn}\right\} &\sim \chi^2_{km(n-1)}.\end{aligned}$$

Using the above and (3.4), and substituting in (2.5), (2.6) and (2.7), we have

$$P\left\{(\chi^2_k/k)\frac{(1+\chi^{-2}_{\nu_b}\nu_b r_n^{(b)} + \chi^{-2}_{\nu_w}\nu_w r_n^{(w)})}{(1 + r_n^{(b)} + r_n^{(w)})} > S_n\right\}. \tag{3.5}$$

The left-hand side of the inequality in (3.5) is approximated as proportional to an F_{k,w_n} distribution by matching the first two moments of each, so that the approximate p-value is $P(\delta F_{k,w_n} > S_n)$, for a proportionality constant δ. Equivalently, the quantity $(1 + \chi^{-2}_{\nu_b}\nu_b r_n^{(b)} + \chi^{-2}_{\nu_w}\nu_w r_n^{(w)})$ is approximated as proportional to an inverse chi-square distributed random variable with degrees of freedom w_n by matching the first two moments of $\eta\chi_w^{-2}$, for proportionality constant η:

$$\begin{aligned}E(\eta\chi_w^{-2}) &= \eta/(w_n - 2)\\ &\approx 1 + \nu_b r_n^{(b)}/(\nu_b - 2) + \nu_w r_n^{(w)}/(\nu_w - 2)\\ E\{(\eta\chi_w^{-2})^2\} &= \eta^2/(w_n-2)(w_n-4)\\ &\approx \frac{2(\nu_w r_n^{(w)})^2}{(\nu_b-2)^2(\nu_w-4)} + \frac{2(\nu_b r_n^{(b)})^2}{(\nu_b-2)^2(\nu_w-4)} + \left(1 + \frac{\nu_b r_n^{(b)}}{\nu_b - 2} + \frac{\nu_w r_n^{(w)}}{\nu_w - 2}\right)^2\end{aligned}$$

Solving these expressions gives the expression for w_n in (3.1) and $\eta = (w_n - 2)(1 + \nu_b r_n^{(b)}/(\nu_b - 2) + \nu_w r_n^{(w)}/(\nu_w - 2))$. Substituting into (3.5), $\delta = (\eta/w_n)/(1 + r_n^{(b)} + r_n^{(w)})$. For sufficiently large ν_b and ν_w, $\delta \approx 1$, so S_n is referred to the F_{k,w_n} distribution.

4. Simulation Studies

In this section, we demonstrate the improved frequentist performance of the test based on w_n over the test based on w_n^* using simulations. We consider only cases where w_n is defined. Shen (2000) shows that tests based on w_n^* when w_n is undefined have good frequentist properties.

For a sample size $s = 1000$, the complete data $\{Y_0, \ldots, Y_{20}\}$ are simulated from independent normal distributions with $E(Y_i) = 0$ for all i, $V(Y_0) = 1$ and $V(Y_i) = 2$ for $i > 0$. For computational simplicity, missingness is simulated by letting $Y_{mis}^{(A)}$ be the first 20% of Y_0 and $Y_{mis}^{(B)}$ be the last 30% of $Y_1, \ldots, Y_{20}$. The partially completed datasets $D_{pcom}^{(i)}, i = 1, \ldots, m$, are generated by drawing values from $f(Y_0|D_{obs})$ using a multivariate normal distribution with an unrestricted covariance matrix. The completed datasets, $\left\{D_{com}^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, n\right\}$ are generated from $f(Y_1, \ldots, Y_{20}|D_{pcom}^{(i)})$, which is a conditional multivariate normal distribution. The number of imputations is varied, with $m \in (2, 5, 10, 20)$ and $n \in (2, 5, 10, 20)$.

The hypothesis tested is $H_0 : Q = 0$, where Q is the vector of coefficients for the regression of Y_0 on $Y_1, \ldots, Y_k$, excluding the intercept, for $k \in (5, 10, 20)$. As this null hypothesis is true in the simulated data, the nominal rejection rate is expected to be close to $100\alpha\%$, for a given significance level α. Table 1 compares the simulated nominal significance levels for 1000 iterations using each combination of m, n, and k, for $\alpha \in (.01, .05, .10)$ using denominator degrees of freedom w_n and w_n^*. The simulated significance levels using w_n are seen to be generally closer to the expected significance levels than when w_n^* is used.

The simulations suggest that the proposed test has appropriate rejection rates when the null hypothesis is true. To get a sense of the power properties, we can turn to the results of Li *et al.* (1991b) and Shen (2000). These tests are derived from similar assumptions and approximations as the test proposed here. Based on extensive simulation studies, Li *et al.* (1991b) report that power curves for their tests are similar to the power curves for Wald-type tests based on the observed data. The greatest losses in power occur when the data deviate substantially from the proportionality assumption. The losses are largest when m is small, and mostly disappear for large m. Shen (2000) reported similar findings, with greatest power loss for small m and n and for large deviations from proportionality. The alternative test proposed here is expected to have similar properties.

The robustness of the test to violations of the proportionality assumptions has been demonstrated for single-stage multiple imputation for missing data in one stage by Li *et al.* (1991a) and for two-stage multiple imputation by Shen (2000). Similar robustness is expected for the alternate reference distribution proposed here.

5. Rates of missing information

Estimates of the fraction of missing information about Q are useful diagnostic tools for assessing how missing data contribute to inferential uncertainty about Q (Schafer, 1997, p. 110). Rubin (1987) addressed estimation of rates of missing information for scalar estimands with single-stage multiple imputation and Harel (2003) addressed asymptotic rates for two-stage multiple imputation. In this section, estimates for single-stage imputation from Rubin (1987) for a finite or infinite number of imputations are reviewed and extended to multivariate estimands, and then further extended to two-stage multiple imputation. Estimates of the rate of missing information in the first stage, the rate of missing information in the second stage, and the overall rate of missing information are given.

Let the subscript *scom* denote quantities derived from D_{scom}, a set of completed datasets for D_{obs} imputed in a single stage, assuming that $Y_{mis}^{(B)}$ is observed, so that the rules of Rubin (1987) apply. When $m \to \infty$, the Fisher information observed for a scalar estimand q is defined to be $(\bar{u}_\infty + b_\infty)^{-1}$, and the total information that would be present if $Y_{mis}^{(A)}$ were also observed is $\bar{u}_\infty^{-1}$; hence the rate of missing information is

$$\gamma^{(A)} = \{\bar{u}_\infty^{-1} - (\bar{u}_\infty + b_\infty)^{-1}\}/\bar{u}_\infty^{-1} = b_\infty(\bar{u}_\infty + b_\infty)^{-1} \tag{5.1}$$

which can be estimated from D_{scom} as $\hat{\gamma}^{(A)} = b_{scom}/(\bar{u}_{scom} + b_{scom})$. Using the posterior distribution $(q|D_{scom}) \sim t_{\nu_m}(\bar{q}_{scom}, T_{scom} = \bar{u}_{scom} + (1 + 1/m)b_{scom})$, the total information about q obtained from D_{scom} when m is finite is given by $(\nu_m + 1)(\nu_m + 3)^{-1}T_{scom}^{-1}$, hence $\gamma^{(A)}$ is estimated by

$$\hat{\gamma}^{(A)} \quad = \quad \{\bar{u}_{scom}^{-1} - (\nu_m + 1)(\nu_m + 3)^{-1}T_{scom}^{-1}\}/\bar{u}_{scom}^{-1} \tag{5.2}$$

where $\nu_m = (m - 1)(1 + 1/r_m)^2$ and $r_m = (1 + 1/m)b_{scom}/\bar{u}_{scom}$. The expression in (5.2) can also be written as

$$\hat{\gamma}^{(A)} = \frac{r_m + 2/(\nu_m + 3)}{1 + r_m}. \tag{5.3}$$

Table 1. Simulated rejection rates for proposed test using w_n and existing test using w_n^*

		w_n (proposed)			w_n^* (existing)		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$\alpha = 0.01$							
	$n = 2$	0.004	0.010	0.009	0.004	0.000	0.000
$m = 2$	$n = 5$	0.002	0.009	0.006	0.003	0.003	0.000
	$n = 10$	0.002	0.012	0.010	0.004	0.003	0.000
	$n = 20$	0.003	0.011	0.010	0.006	0.003	0.000
	$n = 2$	0.010	0.012	0.015	0.004	0.001	0.000
$m = 5$	$n = 5$	0.010	0.011	0.014	0.003	0.001	0.000
	$n = 10$	0.009	0.012	0.013	0.002	0.001	0.000
	$n = 20$	0.009	0.011	0.015	0.005	0.001	0.000
	$n = 2$	0.007	0.010	0.014	0.005	0.004	0.001
$m = 10$	$n = 5$	0.007	0.010	0.014	0.005	0.006	0.001
	$n = 10$	0.006	0.011	0.013	0.003	0.006	0.001
$\alpha = 0.05$							
	$n = 2$	0.014	0.041	0.049	0.019	0.005	0.000
$m = 2$	$n = 5$	0.019	0.051	0.057	0.024	0.010	0.001
	$n = 10$	0.019	0.046	0.065	0.028	0.013	0.001
	$n = 20$	0.020	0.046	0.062	0.025	0.014	0.001
	$n = 2$	0.053	0.060	0.063	0.023	0.014	0.004
$m = 5$	$n = 5$	0.049	0.056	0.072	0.026	0.020	0.010
	$n = 10$	0.050	0.059	0.070	0.029	0.021	0.013
	$n = 20$	0.052	0.056	0.064	0.030	0.022	0.013
	$n = 2$	0.060	0.049	0.069	0.042	0.028	0.023
$m = 10$	$n = 5$	0.057	0.049	0.074	0.040	0.035	0.030
	$n = 10$	0.056	0.053	0.073	0.043	0.037	0.032
$\alpha = 0.10$							
	$n = 2$	0.057	0.091	0.111	0.051	0.014	0.001
$m = 2$	$n = 5$	0.055	0.103	0.116	0.051	0.026	0.002
	$n = 10$	0.051	0.110	0.114	0.046	0.026	0.004
	$n = 20$	0.048	0.110	0.114	0.046	0.027	0.003
	$n = 2$	0.106	0.127	0.131	0.070	0.053	0.027
$m = 5$	$n = 5$	0.107	0.123	0.138	0.076	0.059	0.036
	$n = 10$	0.109	0.127	0.143	0.077	0.057	0.037
	$n = 20$	0.106	0.128	0.136	0.079	0.062	0.039
	$n = 2$	0.099	0.123	0.126	0.091	0.077	0.065
$m = 10$	$n = 5$	0.110	0.124	0.130	0.092	0.083	0.080
	$n = 10$	0.107	0.118	0.128	0.095	0.088	0.081

For multivariate estimands, the posterior of Q generalizes to a multivariate t-distribution, where component q_l of Q has posterior $t_{\nu_m}(\bar{q}_l, T_l)$, $\bar{q}_l$ is the lth component of $\bar{Q}_{scom}$ and T_l is the lth diagonal element of T_{scom}. The degrees of freedom $\nu_m^{(l)}$ for the lth component are $(m-1)(1+1/r_m^{(l)})^2$, where $r_m^{(l)} = (1+1/m)b_{scom}^{(l)}/\bar{u}_{scom}^{(l)}$. As the degrees of freedom $\nu_m^{(l)}$ are the same for each component, we can obtained an improved estimate of ν_m by averaging the $r_m^{(l)}$ across components, yielding

$$r_m = (1+1/m)/k \sum_{l=1}^{k} b_{scom}^{(l)}/\bar{u}_{scom}^{(l)} = (1+1/m)tr(B_{scom}\bar{U}_{scom}^{-1})/k. \tag{5.4}$$

Similarly, under the proportionality assumptions of (3.3), $\gamma^{(A)}$ is the same across components, and hence, to estimate $\gamma^{(A)}$ for multivariate Q, we average the information in Q across components and use (5.3) to estimate $\gamma^{(A)}$, with r_m as defined in (5.4).

With two-stage imputation, D_{scom} is not available, so an estimate of $\gamma^{(A)}$ using D_{com} is needed. To estimate $\gamma^{(A)}$ for Q, note that when using D_{scom}, B_{scom} is an unbiased estimate of B_∞, while when using D_{com}, B provides an unbiased estimate of $B_\infty + \bar{W}_\infty/n$, and $\bar{W}$ is an unbiased estimate of $\bar{W}_\infty$. Thus B_∞ is estimated by $B - \bar{W}/n$ and $\hat{\gamma}^{(A)} = (B - \bar{W}/n)/(\bar{U} + B - \bar{W}/n)$. To estimate $\gamma^{(A)}$ taking into account the finite number of imputations from D_{com}, we use (5.3), replacing r_m with $(1+1/m)tr((B-\bar{W}/n)\bar{U}^{-1})$.

The total fraction of missing information for Q due to both $Y_{mis}^{(A)}$ and $Y_{mis}^{(B)}$ when $m \to \infty$ and $n \to \infty$ is determined similar to (5.2) as $\gamma_{tot} = (B_\infty + \bar{W}_\infty)(\bar{U}_\infty + B_\infty + \bar{W}_\infty)^{-1}$. Since B_∞ is estimated by $B - \bar{W}/n$, an estimate of this fraction is given by $\hat{\gamma}_{tot} = (B + (1-1/n)\bar{W})/(\bar{U} + B + (1-1/n)\bar{W})$. To estimate γ_{tot} when m is finite, note the combining rule for scalar q (Shen, 2000) gives the posterior distribution $(q|D_{com}) \sim t_{\nu_n}(\bar{q}, T_n = \bar{u} + (1+1/m)b + (1-1/n)\bar{w})$, where

$$\nu_n = \left\{ \frac{(r_n^{(b)})^2}{(1+r_n^{(b)}+r_n^{(w)})^2} + \frac{(r_n^{(w)})^2}{(1+r_n^{(b)}+r_n^{(w)})^2} \right\}, \tag{5.5}$$

and $r_n^{(b)}$ and $r_n^{(w)}$ are as defined in (2.6) and (2.7) with $k = 1$. Thus the total information about q in D_{com} is given by $(\nu_n+1)(\nu_n+3)^{-1}T_n^{-1}$, yielding $\hat{\gamma}_{tot} = \{\bar{u}^{-1} - (\nu_n+1)(\nu_n+3)^{-1}T_n^{-1}\}/\bar{u}^{-1}$, which is also written as

$$\hat{\gamma}_{tot} = \frac{2/(\nu_n+3) + r_n^{(b)} + r_n^{(w)}}{1 + r_n^{(b)} + r_n^{(w)}}. \tag{5.6}$$

The assumption of equal fractions of missing information across components in each stage of imputation implies equal fractions of total missing information. Thus, similar to the estimation of $\gamma^{(A)}$ for multivariate Q, we average across components to generalize (5.6) to the multivariate case, using $r_n^{(b)}$ and $r_n^{(w)}$ as defined in (2.6) and (2.7).

An estimate of the fraction of missing information due to $Y_{mis}^{(B)}$ if $Y_{mis}^{(A)}$ were known, assuming an infinite number of imputations, is given by $\gamma^{(B)} = \bar{W}_\infty/(\bar{U}_\infty + \bar{W}_\infty)$. Since $\bar{W}_\infty$ is estimated by $\bar{W}$, an estimate of this fraction is given by $\hat{\gamma}^{(B)} = \bar{W}/(\bar{U} + \bar{W})$. As the denominators of all the rates of missing information considered here are the same, and equal to the total information about Q in the posterior distribution had all the data been observed, $\bar{U}_\infty^{-1}$ or $\bar{U}^{-1}$, estimation of $\gamma^{(B)}$ with finite m can be accomplished by subtraction: $\hat{\gamma}^{(B)} = \hat{\gamma}_{tot} - \hat{\gamma}^{(A)}$.

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 Bayesian logistic regression. *Journal of the American Statistical Association* **86**, 68–78.

Cole, S., Chu, H., and Greenland, S. (2006). Multiple imputation for measurement error correction. *International Journal of Epidemiology* **35**, 1074–1081.

Harel, O. (2003). *Strategies for Data Analysis with Two Types of Missing Values*. Ph.D. thesis, The Pennsylvania State University.

Harel, O. and Schafer, J. (2003). Multiple imputation in two stages. In *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*.

Kim, J. K., Brick, J., Fuller, W., and Kalton, G. (2006). On the bias of the multiple imputation variance estimator in complex sampling. *Journal of the Royal Statistical Society, Ser. B* **68**, 509–521.

Li, K. H., Raghunathan, T. E., Meng, X. L., and Rubin, D. B. (1991a). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica* **1**, 65–92.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991b). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* **86**, 1065–1073.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review* **71**, 593–607.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2007). Multiple imputation when records used for imputation are not used or disseminated for analysis. Tech. rep., Department of Statistical Science, Duke Universtiy.

Reiter, J. P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters* **78**, 15–20.

Reiter, J. P. and Drechsler, J. (2007). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. Tech. rep., Institute for Employment Research (IAB).

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.