Journal of Official Statistics, Vol. 26, No. 2, 2010, pp. 301–315

Tests of Multivariate Hypotheses when using Multiple Imputation for Missing Data and Disclosure Limitation

Satkartar K. Kinney[1] *and Jerome P. Reiter*[2]

Several statistical agencies use, or are considering the use of, multiple imputation to limit the risk of disclosing respondents' identities or sensitive attributes in public use data files. For example, agencies can release partially synthetic datasets, comprising the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. This can be coupled with multiple imputation for missing data in a two-stage imputation approach. First the agency fills in the missing data to generate m completed datasets, then replaces sensitive or identifying values in each completed dataset with n imputed values. Methods for obtaining inferences with the mn datasets have been developed for scalar quantities, but not for multivariate quantities. We present methods for testing multivariate null hypotheses with such datasets. We illustrate the tests using public use files for the Survey of Income and Program Participation that were created with the two-stage imputation approach.

Key words: Confidentiality; disclosure; multiple imputation; significance tests; synthetic data.

1. Introduction

Statistical agencies and other organizations that disseminate data to the public are ethically, practically, and often legally required to protect the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, Rubin (1993) and Little (1993) proposed that agencies utilize multiple imputation approaches. For example, agencies can release the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called partially synthetic datasets (Reiter 2003).

In recent years, statistical agencies have begun to use partially synthetic approaches to create public use data for major surveys. For example, in 2007 the U.S. Census Bureau released a partially synthetic, public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of Social Security benefits information and dozens of other highly sensitive variables (www.sipp.census.gov/sipp/synthdata.html). The U.S. Census Bureau also plans to protect the identities of people in group quarters (e.g., prisons, shelters) in the next release of public use files from the American Communities Survey by replacing demographic data for people at high

[1] National Institute of Statistical Sciences, PO Box 14006, Research Triangle Park, NC 27709, U.S.A. Email: saki@niss.org
[2] Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, U.S.A. Email: jerry@stat.duke.edu
Acknowledgments: This research was supported by the U.S. National Science Foundation grant ITR-0427889.

disclosure risk with imputations. Partially synthetic, public use datasets are in the development stage for the U.S. Census Bureau's Longitudinal Business Database, Longitudinal Employer-Household Dynamics survey, and American Communities Survey veterans and full sample data. Statistical agencies in Canada, Germany (Drechsler et al. 2007), and New Zealand (Graham and Penny 2005) also are investigating the approach. Other applications of partially synthetic data are described by Kennickell (1997), Abowd and Woodcock (2001, 2004), Abowd and Lane (2004), Little et al. (2004), Reiter (2005b), Mitra and Reiter (2006), Reiter and Mitra (2008), An and Little (2007), and Reiter and Raghunathan (2007).

In addition to protecting confidentiality, statistical agencies releasing public use data nearly always have to deal with survey nonresponse. Agencies conveniently can handle the missing data and confidentiality protection simultaneously with a two-stage multiple imputation approach (Reiter 2004). First, the agency uses multiple imputation to fill in the missing data, generating m completed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with n multiple imputations, ultimately releasing mn datasets. The nesting of imputations enables analysts to simply and properly account for the different sources of variability arising from the two types of imputations, which is not straightforward without nesting (Reiter 2004). This two-stage approach was used to create synthetic public use files for the Survey of Income and Program Participation and is being considered for other partially synthetic products at the U.S. Census Bureau.

Methods for obtaining inferences with such two-stage synthetic datasets have been developed for scalar quantities but not for multivariate quantities. Given the importance of the SIPP – it is the largest and most widely used data source on people on public assistance in the U.S. – and the growing interest in using multiple imputation for releasing confidential data, there is a clear need for methodology for obtaining multivariate inferences with such datasets. This article helps address this need by presenting methods for large sample tests of multivariate null hypotheses when multiple imputation is used simultaneously for missing and partially synthetic data.

The remainder of the article is organized as follows. In Section 2, we review the two-stage procedure of Reiter (2004) and extend its distributional theory to multivariate estimands. In Section 3, we use this theory to derive a Wald-like test for multivariate null hypotheses and illustrate the test on the partially synthetic, public use data for the SIPP. In Section 4, we describe an asymptotically equivalent test based on likelihood ratio statistics. Finally, in Section 5, we provide some concluding remarks.

2. Multiple Imputation for Missing Data and Disclosure Limitation

For a finite population of size N, let $I_l = 1$ if unit l is included 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 $s = \sum I_1$. Let X be the $N \times d$ matrix of sampling design variables, e.g., stratum or cluster indicators or size measures. We assume that X is known approximately for the entire population, for example from census records or the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $s \times p$ submatrix of Y for all units with $I_l = 1$, where Y_{obs} is the portion of Y_{inc} that is observed and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let R be an $N \times p$ matrix of indicators such that

$R_{lk} = 1$ if the response for unit l to item k is recorded, and $R_{lk} = 0$ otherwise. The observed data is thus $D_{obs} = (X, Y_{obs}, I, R)$.

To generate the synthetic data, the agency first fills in values for Y_{mis} with draws from the conditional distribution of $(Y_{mis}|D_{obs})$, or approximations of that distribution such as those of Raghunathan et al. (2001). These draws are repeated independently $i = 1, \ldots, m$ times to obtain m completed datasets, $D_{com} = \left\{D_{com}^{(i)} = \left(D_{obs}, Y_{mis}^{(i)}\right), i = 1, \ldots, m\right\}$. Having dealt with the missing data, the agency limits disclosure risks by replacing selected values in each $D_{com}^{(i)}$ with multiple imputations. For each $D_{com}^{(i)}$, imputations are made independently $j = 1, \ldots, n$ times to yield n different partially synthetic data sets. Let $Z_l = 1$ if unit l is selected to have any of its data replaced with synthetic values, and let $Z_l = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \ldots, Z_s)$. Let $Y_{rep}^{(i,j)}$ be all the imputed (replaced) values in the jth synthetic dataset associated with $D_{com}^{(i)}$, and let $Y_{nrep}^{(i)}$ be all unchanged (not replaced) values of $D_{com}^{(i)}$. The $Y_{rep}^{(i,j)}$ are generated from the conditional distribution of $\left(Y_{rep}^{(i,j)}|D_{com}^{(i)}, Z\right)$, or a close approximation of it. Each synthetic dataset, $D_{syn}^{(i,j)}$, then comprises $\left(X, Y_{rep}^{(i,j)}, Y_{nrep}^{(i)}, I, R, Z\right)$. The entire collection of $M = mn$ datasets, $D_{syn} = \left\{D_{syn}^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, n\right\}$, with labels indicating the nests, is released to the public.

We now extend the distributional theory in Reiter (2004) to multivariate quantities. Let Q be a $k \times 1$ vector-valued estimand, such as a vector of regression coefficients. Let $Q^{(i,j)}$ be the estimate of Q computed with $D_{syn}^{(i,j)}$, and let $U^{(i,j)}$ be the estimate of the $k \times k$ covariance matrix of $Q^{(i,j)}$. The following quantities are needed for inferences.

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

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

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{n-1}\sum_{j=1}^{n} (Q^{(i,j)} - \bar{Q}^{(i)})(Q^{(i,j)} - \bar{Q}^{(i)})' = \frac{1}{m}\sum_{i=1}^{m} W^{(i)}$$

$$B = \frac{1}{m-1}\sum_{i=1}^{m} (\bar{Q}^{(i)} - \bar{Q})(\bar{Q}^{(i)} - \bar{Q})'$$

We also define $B_\infty = \lim B \text{ as } m \to \infty \text{ and } n \to \infty$, $W_\infty^{(i)} = \lim W^{(i)} \text{ as } n \to \infty$; and, $\bar{W}_\infty = \sum_{i=1}^{m} W_\infty^{(i)}/m$. All posterior distributions presented in this and subsequent sections are based on diffuse prior distributions.

To begin, we utilize the assumptions described by Rubin (1987, Chapter 3) for multiple imputation for missing data. Let $Q_{com}^{(i)}$ be the estimate of Q that would be obtained from $D_{com}^{(i)}$ prior to replacement of confidential values, and let $\bar{Q}_{com} = \sum_{i=1}^{m} Q_{com}^{(i)}/m$. We then can write Rubin's (1987) well-known result as

$$(Q|D_{com}, B_\infty, \bar{W}_\infty) \sim N(\bar{Q}_{com}, \bar{U} + (1 + 1/m)B_\infty) \tag{1}$$

An implicit assumption here is that each $U^{(i,j)}$ has sufficiently low variability so that $U^{(i,j)} \approx U_{com}^{(i)}$, where $U_{com}^{(i)}$ is the variance estimate of $Q_{com}^{(i)}$ computed from $D_{com}^{(i)}$. Similarly, we assume that the $U_{com}^{(i)} \approx U$, and hence $\bar{U} \approx U$, where U is the variance that

would be obtained from $D_{inc} = (X, Y_{inc}, I)$, i.e., if all the data were observed. These are typical assumptions in multiple imputation, motivated by the fact that posterior variances generally have lower order variability than posterior means (Rubin 1987, p. 89).

Following Reiter (2004), we assume the sampling distributions $Q^{(i,j)} \sim N\left(Q_{com}^{(i)}, W_{\infty}^{(i)}\right)$ for all (i, j), so that

$$\left(Q_{com}^{(i)} | D_{syn}, B_{\infty}, W_{\infty}^{(i)}\right) \sim N\left(\bar{Q}^{(i)}, W_{\infty}^{(i)}/n\right) \tag{2}$$

Integrating (1) and (2) with respect to the collection of $Q_{com}^{(i)}$, we have

$$\left(Q | D_{syn}, B_{\infty}, \bar{W}_{\infty}\right) \sim N\left(\bar{Q}, T_{\infty}\right) \tag{3}$$

where $T_{\infty} = \bar{U} + (1 + 1/m)B_{\infty} + \bar{W}_{\infty}/(mn)$. We note that the fractional increase in the variance of Q due to missing data is $(1 + 1/m)B_{\infty}\bar{U}^{-1}$ and due to replacement data is $\{\bar{W}_{\infty}/(mn)\}\bar{U}^{-1}$.

In practice, B_{∞} and $\bar{W}_{\infty}$ are not known and must be integrated out of (3). To do so, we utilize the sampling distributions for $Q_{com}^{(i)}$ from Rubin (1987), $Q_{com}^{(i)} \sim N(Q_{obs}, B_{\infty})$ for all i. Here, Q_{obs} is the estimate of Q that would be obtained from D_{obs}. Combining these distributions with the sampling distribution underlying (2), we have

$$\left(\bar{Q}^{(i)} | D_{obs}, B_{\infty}, W_{\infty}^{(i)}\right) \sim N\left(Q_{obs}, B_{\infty} + W_{\infty}^{(i)}/n\right) \tag{4}$$

Using the sampling distribution of $Q^{(i,j)}$, we have

$$\left\{W^{(i)}\left(W_{\infty}^{(i)}\right)^{-1} \middle| D_{syn}\right\} \sim Wi(n-1, I) \tag{5}$$

where $Wi(n-1, I)$ is a Wishart distribution with $(n-1)$ degrees of freedom and scale matrix I.

Finally, from (4) and (5) and the simplifying assumption that $W_{\infty}^{(i)} = \bar{W}_{\infty}$ for all i, we have

$$\{B(B_{\infty} + \bar{W}_{\infty}/n)^{-1} | D_{syn}, \bar{W}_{\infty}\} \sim Wi(m-1, I) \tag{6}$$

$$\{\bar{W}(\bar{W}_{\infty})^{-1} | D_{syn}\} \sim Wi(m(n-1), I) \tag{7}$$

For sufficiently large s, m, and n, we can replace B_{∞} and each $\bar{W}_{\infty}$ with their approximate expected values, resulting in the variance estimate $T = (1 + 1/m)B - (1/n)\bar{W} + \bar{U}$. Analysts can base inferences for Q on the distribution,

$$(Q - \bar{Q}) \sim N(0, T) \tag{8}$$

The fractional increase in variance due to missing data is estimated from D_{syn} to be $(1 + 1/m)(B - \bar{W}/n)\bar{U}^{-1}$. The estimated fractional increase due to replacement data is $\{\bar{W}/(mn)\}\bar{U}^{-1}$. For each of these, the average fractional increase across components of Q equals the average of the diagonal elements of these matrices.

3. Wald-type Tests

Using the M released datasets, an analyst seeks to test the null hypothesis $Q = Q_0$, for example to test if k regression coefficients equal zero. In this section, we first argue and demonstrate that the natural test based on the Wald test statistic for (8) can be poorly calibrated. We then derive an alternative test that tends to have better properties.

3.1. Poor Properties of Test Based on Wald Test Statistic

Given the normal approximation for inferences about Q in (8), it may appear reasonable to use the test statistic, $\Delta = (Q_0 - \bar{Q})'T^{-1}(Q_0 - \bar{Q})$, and the p-value equal to $pr(\chi_k^2 > \Delta)$, where χ_k^2 is a chi-squared random variable with k degrees of freedom. However, this test is unreliable when k is large and m and n are moderate, as is frequently the case, because B or $\bar{W}$ can have large variability. Estimating B or $\bar{W}$ in such cases is akin to estimating a covariance matrix using few observations compared to the number of dimensions.

We can illustrate the poor properties of tests based on Δ with simulation studies. For sample size $s = 1{,}000$, we simulate the complete data, $\{Y_0, Y_1, \ldots, Y_{20}\}$, from independent normal distributions with $E(Y_i) = 0$ for all i, $var(Y_0) = 1$, and $var(Y_i) = 2$ for $i > 0$. To simulate missing data, for computational simplicity we make 30% of the observations have $\{Y_1, \ldots, Y_{20}\}$ missing completely at random and Y_0 always fully observed. We obtain the set of completed datasets, D_{com}, by drawing values of the missing data from $f(Y_1, \ldots, Y_{20}|D_{obs})$, using a multivariate normal distribution with an unrestricted covariance matrix. To simulate partial synthesis, we replace all values of Y_0. The replacement imputations for each $D_{syn}^{(i,j)}$ are drawn independently from $f(Y_0|D_{com}^{(i)})$. We vary the number of imputations according to $m \in (4,8)$ and $n \in (2,4,8)$, which are in the range of values likely to be used by agencies releasing data. For example, the SIPP synthetic data use ($m = 4$, $n = 4$).

We test the null hypothesis $Q = 0$, where Q is the vector of coefficients for the regression of Y_0 on $Y_1, \ldots, Y_k$, excluding the intercept, for $k \in (5, 10, 20)$. Table 1 summarizes the simulated rejection rates of the test based on $pr(\chi_k^2 > \Delta)$. Results are based on 10,000 runs of the simulation for each combination of m, n, and k, for significance levels $\alpha \in (1\%, 5\%, 10\%)$. The rejection rates for the test far exceed the nominal α levels, often by so much as to make the test essentially useless. This problem is alleviated by making m and n excessively large. In the simulation scenario just described, setting m and n to 50 yielded rejection rates much closer to the desired levels; however, doing so in practice is impractical. In addition to computational and analytic burden, releasing so many datasets can result in increased risk of disclosure.

Another problem with the test statistic Δ is that the variance T can have negative diagonal elements, which can result in negative values of Δ. This is most likely to occur for large values of k when n is small. This occurred in a simulation test about 20% of the time

Table 1. Simulated rejection rates in percentages for significance levels α using $pr(\chi_k^2 > \Delta)$

	$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 10\%$		
k Value:	5	10	20	5	10	20	5	10	20
$m = 4$									
$n = 2$	10.0	11.3	21.8	16.0	21.3	38.6	20.7	28.2	49.1
$n = 4$	26.0	32.4	33.3	37.1	40.9	40.0	43.9	45.6	43.7
$n = 8$	8.0	21.1	54.1	19.4	37.7	72.0	27.6	48.1	79.8
$m = 8$									
$n = 2$	11.8	10.7	10.2	17.2	15.1	16.8	21.0	18.7	22.8
$n = 4$	11.7	36.0	39.3	22.2	48.0	45.9	30.0	55.1	49.8

with small values of n, with the percentage fading to zero when $m = 4$ and $n = 8$. An ad-hoc adjustment, used in Table 1, is to replace T with $\bar{U} + (1 + 1/m)B$ (Reiter 2008).

3.2. An Improved Wald-type Test

Rubin (1987), Li et al. (1991a,b), and Reiter (2005a) identified similar issues in multivariate testing in single-stage multiple imputation procedures. To mitigate the effects of variability, they reduce the number of unknown parameters in B_∞ (there is no $\bar{W}_\infty$ in one stage imputation) and derive an alternative to the natural Wald test statistic. We adapt this strategy for two-stage partially synthetic data to propose an improved Wald-like test. Our adaptation deals with both variance estimates (B and $\bar{W}$) that can make Δ unstable.

We first present the test statistic and its reference distribution for the improved Wald-like test, followed by the derivation. The statistic is

$$S = (Q_0 - \bar{Q})' U^{-1} (Q_0 - \bar{Q}) / \{k(1 + r^{(b)} - r^{(w)})\}$$

where

$$r^{(b)} = (1 + 1/m)\text{tr}(B\bar{U}^{-1})/k \tag{9}$$

$$r^{(w)} = (1/n)\text{tr}(\bar{W}\bar{U}^{-1})/k \tag{10}$$

The reference distribution is approximated by an F_{k,w_s} distribution where

$$w_s = 4 + \frac{\{1 + ((r^{(b)}v_b)/(v_b - 2)) - ((r^{(w)}v_w)/(v_w - 2))\}^2}{((r^{(b)}v_b)^2)/((v_b - 2)^2(v_b - 4)) + ((r^{(w)}v_w)^2)/((v_w - 2)^2(v_w - 4))} \tag{11}$$

for $v_b > 4$ and $v_w > 4$, and $v_b = k(m - 1)$ and $v_w = km(n - 1)$. We provide an alternate degrees of freedom for the special case where $v_b \leq 4$ or $v_w \leq 4$ at the end of this section. The approximate p-value for testing $Q = Q_0$ is given by $pr(F_{k,W_s} > S)$.

3.2.1. Derivation of Test

Conditional on T_∞ and using standard multivariate theory with (3), the p-value for testing $Q = Q_0$ is $pr\left(\chi^2_k > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q})\right)$. Since T_∞ is unknown, we average this probability over the distribution of T_∞, or equivalently over the distributions of $(B_\infty | D_{syn}, \bar{W}_\infty)$ and $(\bar{W}_\infty | D_{syn})$ in (6) and (7). Averaging over the variance parameters, the p-value equals

$$\int pr\left\{\chi^2_k > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) | D_{syn}, B_\infty, \bar{W}_\infty\right\}$$

$$\times pr(B_\infty | D_{syn}, \bar{W}_\infty) pr(\bar{W}_\infty | D_{syn}) dB_\infty d\bar{W}_\infty$$

This integral can be evaluated numerically, but it is desirable to have a simple, closed-form approximation for analysts of public use data, who may not have the skills or desire to perform the numerical integration.

For the approximation, we set $B_\infty = r_\infty^{(b)} \bar{U}_\infty$ and $\bar{W}_\infty = r_\infty^{(w)} \bar{U}_\infty$, where $r_\infty^{(w)}$ and $r_\infty^{(b)}$ are scalar quantities not assumed to be equal. These equations are true if (i) the fractions of missing information are equal for all components of Q, and (ii) the fractions of replaced information are equal for all components of Q. These conditions do not strictly hold in all

surveys; however, in practice, rates of missing and replaced information often do not vary substantially by variable. In such cases, the stabilization in the estimate of T_∞ resulting from these approximations can lead to tests with better properties than tests based on Δ. We illustrate this using simulations in Section 3.2.2.

Under these conditions, $T_\infty = \bar{U}_\infty\{1 + (1 + 1/m)r_\infty^{(b)} + r_\infty^{(w)}/mn\}$, so that the number of parameters to be estimated for each of B_∞ and $\bar{W}_\infty$ is reduced from $k(k+1)/2$ to 1, thereby stabilizing the estimation of T_∞. Assuming $\bar{U} \approx \bar{U}_\infty$, the p-value becomes

$$\begin{aligned}
&\int pr\left\{\chi_k^2 > \frac{(Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})}{1 + (1 + 1/m)r_\infty^{(b)} + r_\infty^{(w)}/(mn)} \middle| D_{syn}, r_\infty^{(b)}, r_\infty^{(w)}\right\} \\
&\quad \times pr\left(r_\infty^{(b)} | D_{syn}, r_\infty^{(w)}\right) pr\left(r_\infty^{(w)} | D_{syn}\right) dr_\infty^{(b)} dr_\infty^{(w)} \\
&\quad = \int pr\left\{\left(\chi_k^2/k\right)\frac{1 + (1 + 1/m)r_\infty^{(b)} + r_\infty^{(w)}/(mn)}{(1 + r^{(b)} + r^{(w)})} > S \middle| D_{syn}, r_\infty^{(b)}, r_\infty^{(w)}\right\} \\
&\quad \times pr\left(r_\infty^{(b)} | D_{syn}, r_\infty^{(w)}\right) pr\left(r_\infty^{(w)} | D_{syn}\right) dr_\infty^{(b)} dr_\infty^{(w)}
\end{aligned} \tag{12}$$

The posterior distributions of $\left(r_\infty^{(b)} | D_{syn}, r_\infty^{(w)}\right)$ and of $\left(r_\infty^{(w)} | D_{syn}\right)$ can be obtained from (6) and (7). Applying standard multivariate normal theory, we have

$$\left\{\frac{k(m-1)\mathrm{tr}(B\bar{U}^{-1})/k}{r_\infty^{(b)} + r_\infty^{(w)}/n} \middle| D_{syn}, r_\infty^{(w)}\right\} \sim \chi^2_{k(m-1)} \tag{13}$$

$$\left\{\frac{km(n-1)\mathrm{tr}(\bar{W}\bar{U}^{-1})/k}{r_\infty^{(w)}} \middle| D_{syn}\right\} \sim \chi^2_{km(n-1)} \tag{14}$$

Substituting (13) and (14) into (12), after some algebra we have

$$pr\left\{\left(\chi_k^2/k\right)\frac{1 + v_b r^{(b)}/\chi^2_{v_b} - v_w r^{(w)}/\chi^2_{v_w}}{1 + r^{(b)} - r^{(w)}} > S\right\} \tag{15}$$

We approximate the random variable in (15) as proportional to an F-distributed random variable, F_{k,w_s}, so that the p-value is $pr(\delta F_{k,w_s} > S)$. The approximation is obtained by matching the first two moments of $\delta F_{k,w_s}$ to those of the left-hand side of the inequality in (15). Equivalently, we approximate $\left(1 + \chi^{-2}_{v_b} v_b r^{(b)} - \chi^{-2}_{v_w} v_w r^{(w)}\right)$ as proportional to an inverse chi-square distributed random variable with degrees of freedom w_s by matching the first two moments to the distribution $\eta\chi^{-2}_{w_s}$. Using iterated expectations and variances, we have

$$E\left(\eta\chi^{-2}_{w_s}\right) = \frac{\eta}{(w_s - 2)} \approx 1 + \frac{v_b r^{(b)}}{v_b - 2} - \frac{v_w r^{(w)}}{v_w - 2}$$

and

$$E\left\{\left(\eta\chi^{-2}_{w_s}\right)^2\right\} = \frac{\eta^2}{(w_s-2)(w_s-4)}$$
$$\approx \frac{2(v_w r^{(w)})^2}{(v_w-2)^2(v_w-4)} + \frac{2(v_b r^{(b)})^2}{(v_b-2)^2(v_b-4)} + \left\{E\left(\eta\chi^{-2}_{w_s}\right)\right\}^2$$

Solving yields the expression in (11) for w_s and $\delta = \{(w_s-2)/w_s\}\{1 + v_b r^{(b)}/(v_b-2) - v_w r^{(w)}/(v_w-2)\}/(1 + r^{(b)} - r^{(w)})$. When v_b and v_w are sufficiently large, $\delta \approx 1$, and the approximate p-value is $pr(F_{k,w_s} > S)$.

It is possible that $v_b \leq 4$ or $v_w \leq 4$, in which case w_s is not defined. This can occur for small k when $m = 2$, a choice for m that is not recommended due to the potentially high probability that estimated variances for scalar quantities are less than zero (Reiter 2008). Nonetheless, if analysts find that $v_b \leq 4$, we suggest the alternate denominator degrees of freedom,

$$w_s^* = \left\{\frac{(r^{(b)})^2}{v_b(1 + r^{(b)} - r^{(w)})^2} + \frac{(r^{(w)})^2}{v_w(1 + r^{(b)} - r^{(w)})^2}\right\}^{-1} \tag{16}$$

This is a generalization of the degrees of freedom used in the t-distribution of Reiter (2004) for inferences for scalar Q. Details of its derivation can be found in Kinney (2007).

3.2.2. Illustration of Improved Performance

To illustrate the improved performance of this test, we repeat the simulations of Section 3.1. In this simulation scenario, the fractions of missing information on each component of Q are equal, as are the fractions of replaced information for each component of Q. Table 2 displays the simulated significance levels for the null hypothesis $Q = 0$, where Q is the vector of coefficients for the regression of Y_0 on $Y_1, \ldots, Y_k$, excluding the intercept, for $k \in (5, 10, 20)$. The simulated levels are much closer to the α-levels than those based on tests with Δ (displayed in Table 1). Additionally, S was observed to be positive in all 10,000 runs for each scenario.

In some applications of partially synthetic data, including SIPP, several variables are replaced in entirety while others are left unchanged. In such cases, when Q involves both

Table 2. Simulated rejection rates in percentages for significance levels α using $pr(F_{k,w_s} > S)$

	$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 10\%$		
k Value:	5	10	20	5	10	20	5	10	20
$m = 4$									
$n = 2$	0.1	0.3	0.6	1.8	3.0	4.2	5.1	7.3	9.2
$n = 4$	0.7	1.0	1.2	3.9	5.1	5.3	8.7	10.2	10.7
$n = 8$	1.0	1.2	1.0	4.7	5.0	5.4	9.4	10.3	10.5
$m = 8$									
$n = 2$	0.4	0.7	0.9	3.1	4.2	5.1	7.2	9.1	10.3
$n = 4$	0.8	1.2	1.2	5.1	5.3	5.6	10.3	10.7	10.9

replaced and unreplaced variables the fractions of replaced information are not equal across components of Q, i.e., $\bar{W}_\infty \neq r_\infty^{(w)} U_\infty$. Additionally, it is commonly the case that the fraction of missing information is not equal across components of Q and so $\bar{W}_\infty \neq r_\infty^{(w)} U_\infty$. To gain insight on the performance of the procedures when the proportionality conditions do not hold, we can turn to the literature on significance testing in multiple imputation for missing data only. Using simulations, Li et al. (1991a) show that tests based on the condition, $B_\infty = r_\infty U_\infty$, are robust in cases of practical interest when that condition does not hold. That is, tests based on the proportionality condition are better calibrated than those based on the corresponding natural Wald test. This suggests that tests based on S still should outperform those based on Δ in cases of practical interest.

We illustrate the robustness of the proposed test to violations of both proportionality assumptions by modifying the simulation scenario of Table 2. We create unequal fractions of missing information by letting $Y_0, \ldots, Y_{10}$ be completely observed and $Y_{11}, \ldots, Y_{20}$ be 30% missing. We create unequal fractions of replaced information by replacing all values in $Y_0, \ldots, Y_{10}$ with imputations in the second stage and leaving $Y_{11}, \ldots, Y_{20}$ unchanged in the second stage of imputation. We set $k = 20$ and test $Q = 0$, where Q is the vector of coefficients from the regression of Y_0 on $Y_1, \ldots, Y_{20}$. Table 3 gives the simulated rejection rates over 1,000 iterations. These are slightly conservative but close the desired levels. While not shown, tests based on Δ continue to be very poorly calibrated.

3.3. Application With SIPP Public Use Data

We now illustrate the application of the Wald-like test using the partially synthetic data from the SIPP public use files. We first provide a brief overview of the SIPP, followed by the application.

The SIPP is a continuous series of national panels designed to collect data on income, labor force information, participation and eligibility for governmental assistance programs, and general demographic characteristics for individuals on public assistance. It can be used for both longitudinal and cross-sectional analyses, including assessments of the effectiveness and impacts of changes in public assistance programs, the distributions of wealth across different demographic groups, and the factors that affect changes in household and family structures (www.sipp.census.gov/sipp/analytic.html). The national panels range in size from approximately 14,000 to 36,700 interviewed households and last from two and a half to four years. Households are selected in a multistage, stratified

Table 3. Simulated rejection rates in percentages for significance levels α with $k = 20$ using $pr(F_{k,w_s} > S)$ when the proportionality assumptions are not valid

	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
$m = 4$			
$n = 2$	1.1	5.3	10.4
$n = 4$	1.2	5.8	10.6
$n = 8$	1.5	5.3	10.4
$m = 8$			
$n = 2$	1.2	5.7	10.9
$n = 4$	1.3	5.5	10.5

sampling design (www.sipp.census.gov/sipp/overview.html). Analysts can download de-identified data from the SIPP website.

In 2001, the U.S. Census Bureau, the Internal Revenue Service, and the Social Security Administration decided to supplement the information on SIPP panels from 1990–1996 with detailed earnings and Social Security benefits histories. Linking these data allows researchers to study retirement and disability programs and their interactions with other public assistance programs. Because of the highly sensitive nature of these supplemental data, the three agencies agreed to release a version of the linked data only if sensitive and identifying information was synthesized. In the end, the agencies determined that it was necessary to synthesize all but four out of over six hundred variables in the linked data.

The linked data (before synthesis) also contains a large number of missing values, some due to the panel structure and others to survey or administrative database nonresponse. Therefore, the team developing the synthesis used the two stage approach to creating synthetic data. First, they generated $m = 4$ completed datasets using a combination of sequential regression multivariate imputation (Raghunathan et al. 2001) and Bayesian bootstraps (Rubin 1981). Then, for each completed dataset, they generated $n = 4$ synthetic copies by replacing all values of the sensitive variables. The synthesis was done using sequential regression multivariate imputation and a kernel density regression technique developed by Woodcock and Benedetto (2006). These $M = 16$ datasets are released to the public and available for downloading on the SIPP website. For details of the synthesis procedure, see Abowd et al. (2006).

Abowd et al. (2006) report on several estimands and regressions of interest using SIPP data, providing univariate confidence intervals computed using the methods of Reiter (2004), and comparing with estimates from the completed data prior to synthesis. As a practical example, we illustrate a multivariate test using the regression of log total family income in 1999 against number of children, year of birth, indicators for gender, black, Hispanic, foreign born, and disabled, and categorical variables indicating education level, marital status, and type of benefits received. The multivariate test was applied to see if a 4-level categorical variable indicating industry type should be included in the regression model. In this case, test statistic Δ was negative, yielding a p-value of 1 when compared to a χ^2_3-distribution; when adjusted as in Section 3.1, the test statistic was 203, yielding a p-value of 10. On the other hand, the test statistic S had a value of 32, which yielded a p-value of .0004 when compared to the $F_{3,w}$-distribution, where $w = 8.89$, suggesting that the industry variable is a significant predictor of income.

4. Test Based on Likelihood Ratio Statistics

The Wald-like test requires access to all elements of the $U^{(i,j)}$ matrices. This may be cumbersome when the dimension of $U^{(i,j)}$ is large. We now present a test based on the set of log-likelihood ratio statistics from the completed datasets. This test is similar in spirit to those developed by Meng and Rubin (1992) and Shen (2000) for multiple imputation for missing data only and by Reiter (2005a) for multiple imputation for synthetic data only. As before, we first present the test before outlining its derivation.

Following the notation in Schafer (1997), let ψ be the vector of parameters in the analyst's model. Let $\hat{\psi}_0^{(i,j)}$ and $\hat{\psi}^{(i,j)}$ be the maximum likelihood estimates of

Q computed with $D_{syn}^{(i,j)}$ under the null and alternative hypotheses, respectively. Let $\bar{\psi}^{(i)} = \sum_{j=1}^{n} \hat{\psi}^{(i,j)}/n$; $\bar{\psi}_0^{(i)} = \sum_{j=1}^{n} \hat{\psi}_0^{(i,j)}/n$; $\bar{\psi} = \sum_{i=1}^{m} \hat{\psi}^{(i)}/m$; and, $\bar{\psi}_0 = \sum_{i=1}^{m} \bar{\psi}_0^{(i)}/m$. We write the log-likelihood ratio statistic evaluated at any two values a and b for any dataset $D_{syn}^{(i,j)}$ as $d'\left(a, b|D_{syn}^{(i,j)}\right) = 2\log f\left(D_{syn}^{(i,j)}|a\right) - 2\log f\left(D_{syn}^{(i,j)}|b\right)$. The test statistic is

$$\tilde{S} = \bar{L}/\{k(1 + \tilde{r}^{(b)} - \tilde{r}^{(w)})\} \tag{17}$$

where

$$\tilde{r}^{(b)} = \{(m+1)(\bar{L}_m - \bar{L})\}/\{k(m-1)\}$$

$$\tilde{r}^{(w)} = (\bar{l} - \bar{L}_m)/\{k(n-1)\}$$

and

$$\bar{L} = \sum_{i=1}^{m}\sum_{j=1}^{n} d'\left(\bar{\psi}_0, \bar{\psi}|D_{syn}^{(i,j)}\right)/(mn)$$

$$\bar{L}_m = \sum_{i=1}^{m}\sum_{j=1}^{n} d'\left(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)}|D_{syn}^{(i,j)}\right)/(mn)$$

$$\bar{l} = \sum_{i=1}^{m}\sum_{j=1}^{n} d'\left(\hat{\psi}_0^{(i,j)}, \hat{\psi}^{(i,j)}|D_{syn}^{(i,j)}\right)/(mn)$$

The reference distribution for $\tilde{S}$ is an F-distribution with k degrees of freedom in the numerator and $\tilde{w}_s$ degrees of freedom in the denominator, where $\tilde{w}_s$ is the expression in (11) with the terms $r^{(b)}$ and $r^{(w)}$ replaced by $\tilde{r}^{(b)}$ and $\tilde{r}^{(w)}$. When $v_b \leq 4$ or $v_w \leq 4$, we use the denominator degrees of freedom in (16), substituting in $\tilde{r}^{(b)}$ and $\tilde{r}^{(w)}$ as above.

The derivation parallels the strategy of Meng and Rubin (1992), namely (i) find a statistic asymptotically equivalent to S based only on the Wald statistics from each synthetic dataset; (ii) use the asymptotic equivalence of Wald and log-likelihood ratio test statistics for individual datasets to define the test statistic $\tilde{S}$; and, (iii) find a reference F-distribution as in the Wald tests.

To begin, let $d(Q^{(i,j)}, U^{(i,j)}) = (Q^{(i,j)} - Q_0)' U^{(i,j)-1}(Q^{(i,j)} - Q_0)$ for all (i, j). Because of the asymptotic equivalence of Wald and log-likelihood ratio test statistics, each $d(Q^{(i,j)}, U^{(i,j)})$ is asymptotically equivalent to its corresponding $d'\left(\hat{\psi}_0^{(i,j)}, \hat{\psi}^{(i,j)}|D_{syn}^{(i,j)}\right)$. Furthermore, because of the low-order variability in the $U^{(i,j)}$, we can interchange the $U^{(i,j)}$ with $\bar{U}$ in any of $d(Q^{(i,j)}, U^{(i,j)})$, $d(\bar{Q}^{(i)}, U^{(i,j)})$, or $d(\bar{Q}, U^{(i,j)})$.

Let $\bar{d} = \sum_{i=1}^{m}\sum_{j=1}^{n} d(Q^{(i,j)}, U^{(i,j)})/(mn)$; let $\bar{d}^{(i)} = \sum_{j=1}^{n} d(\bar{Q}^{(i)}, U^{(i,j)})/n$; and, let $\hat{d} = \sum_{i=1}^{m}\sum_{j=1}^{n} d(\bar{Q}, U^{(i,j)})/(mn)$. Then S is equivalent to

$$S^* = \frac{(\bar{d}/k) - (n-1)r^{(w)} - (m-1)r^{(b)}/(m+1)}{1 + r^{(b)} - r^{(w)}} \tag{18}$$

where $r^{(b)}$ and $r^{(w)}$ are defined in (9) and (10). To show this, we assume without loss of generality that $Q_0 = 0$ and $\bar{U}$ is a $k \times k$ identity matrix, as in Rubin (1987, p. 100). Then,

$S = \bar{Q}'\bar{Q}/\{k(1 + r^{(b)} - r^{(w)})\}$ and, using a sums-of-squares decomposition,

$$\bar{d} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(Q^{(i,j)} - \bar{Q}^{(i)})'(Q^{(i,j)} - \bar{Q}^{(i)})$$

$$+ \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{Q}^{(i)} - \bar{Q})'(\bar{Q}^{(i)} - \bar{Q}) + \bar{Q}'\bar{Q}$$

$$= k(n-1)r^{(w)} + \frac{k(m-1)}{m+1}r^{(b)} + \bar{Q}'\bar{Q}$$

Substituting the above expression into (18) yields S.

Computing $r^{(b)}$ and $r^{(w)}$ requires access to $\bar{U}$, which we do not want these tests to depend on. Expressions that rely only on Wald statistics are obtained by using sums-of-squares decompositions. Under the canonical conditions, and without loss of generality, for $r^{(b)}$ we have

$$r^{(b)} = \frac{(m+1)}{km(m-1)}\sum_{i=1}^{m}(\bar{Q}^{(i)} - \bar{Q})'(\bar{Q}^{(i)} - \bar{Q})$$

$$= \frac{(m+1)}{km(m-1)}\left\{\sum_{i=1}^{m}(\bar{Q}^{(i)\prime}\bar{Q}^{(i)}) - m\bar{Q}'\bar{Q}\right\} \quad \approx \frac{(m+1)}{k(m-1)}\left(\sum_{i=1}^{m}\bar{d}^{(i)}/m - \hat{d}\right) = r_w^{(b)}$$

since $\sum_{i=1}^{m}\bar{d}^{(i)}/m$ is asymptotically equivalent to $\sum_{i=1}^{m}(\bar{Q}^{(i)\prime}\bar{Q}^{(i)})$, and $\hat{d}$ is asymptotically equivalent to $\bar{Q}'\bar{Q}$. For $r^{(w)}$, we have

$$r^{(w)} = \frac{1}{kmn(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(Q^{(i,j)} - \bar{Q}^{(i)})'(Q^{(i,j)} - \bar{Q}^{(i)})$$

$$= \frac{1}{kmn(n-1)}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}(Q^{(i,j)\prime}Q^{(i,j)}) - n\sum_{i=1}^{n}(\bar{Q}^{(i)\prime}\bar{Q}^{(i)})\right\}$$

$$\approx k(n-1)\left(\bar{d} - \sum_{i=1}^{m}\bar{d}^{(i)}/m\right) = r_w^{(w)}$$

Using $\hat{d}$ to approximate the numerator of S, and $r_w^{(b)}$ and $r_w^{(w)}$ to approximate $r^{(b)}$ and $r^{(w)}$ in the denominator of S, we obtain the asymptotically equivalent statistic S^*.

We next utilize the asymptotic equivalence between the Wald statistics and the log-likelihood ratio statistic to show that $\tilde{S}$ in (17) is asymptotically equivalent to S^*. The equivalence of $\bar{l}$ and $\bar{d}$ follows directly from the asymptotic equivalence of the $d(Q^{(i,j)},U_{ij})$ and their corresponding $d'\left(\hat{\psi}^{(i,j)}, \hat{\psi}_0^{(i,j)}|D_{syn}^{(i,j)}\right)$. The equivalence of $\bar{L}$ and $\hat{d}$, and of $\bar{d}_m = \sum_{i=1}^{m}\bar{d}^{(i)}/m$ and $\bar{L}_m$, is more subtle. Using arguments similar to those of Meng and Rubin (1992) and Shen (2000), for quadratic complete-data log-likelihood functions, we have

$$d'(\bar{\psi}_0, \bar{\psi}|D_{syn}^{(i,j)}) \approx d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}, U^{(i,j)})$$

$$d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)}|D_{syn}^{(i,j)}) \approx d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}^{(i)}, U^{(i,j)})$$

Thus, we have

$$\begin{aligned}
\bar{L} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d'\left(\bar{\psi}_0, \bar{\psi} | D_{syn}^{(i,j)}\right) \\
&\approx \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \{d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}, U^{(i,j)})\} \\
&\approx \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \{d(Q^{(i,j)} - \bar{U}) - d(Q^{(i,j)} - \bar{Q}, \bar{U})\} \\
&\approx d(\bar{Q}, \bar{U})\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d(\bar{Q}, U^{(i,j)}) = \hat{d}
\end{aligned}$$

Similar reasoning shows that $\sum_{i=1}^{m}\bar{d}^{(i)}/m$ is asymptotically equivalent to $\bar{L}_m$. Thus, we can replace $\bar{l}$ with $\bar{d}$, $\bar{L}$ with $\hat{d}$, and $\bar{L}_m$ with $\bar{d}_m$ to obtain the test statistic $\tilde{S}$ and reference F-distribution.

5. Concluding Remarks

The simulations suggest that the improved Wald-like test provides appropriate rejection rates when the null hypothesis is true. To get a sense of the power properties of these tests, we can turn to the results of Li et al. (1991b), who examined the power properties of large sample significance tests for multiple imputation of missing data only. These tests are derived from similar assumptions and approximations as the Wald-like 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 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 for nested imputation, with greatest power loss for small m and n and for large deviations from proportionality. The tests proposed here are expected to have similar properties, though further study is needed.

Popular software packages contain routines for obtaining confidence intervals for scalar quantities and p-values for multicomponent tests from multiply-imputed datasets. These routines can be easily modified to perform the tests proposed here.

As resources available to malicious data users continue to expand, the alterations needed to protect data with traditional disclosure limitation techniques – such as swapping, adding noise, or microaggregation – may become so extreme that, for many analyses, the released data are no longer useful. Synthetic data, on the other hand, has the potential to enable data dissemination while preserving data utility. The methods in this article enable analysts of multiply-imputed, partially synthetic public-use data to obtain closer to nominal levels when testing multicomponent null hypotheses than previously possible, thereby increasing the utility of synthetic data approaches.

6. References

Abowd, J.M. and Lane, J.I. (2004). New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. In Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra (eds). New York: Springer-Verlag, 282–289.

Abowd, J.M., Stinson, M.H., and Benedetto, G.L. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical Report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program.

Abowd, J.M. and Woodcock, S.D. (2001). Disclosure Limitation in Longitudinal Linked Data. In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). Amsterdam: North-Holland, 215–277.

Abowd, J.M. and Woodcock, S.D. (2004). Multiply-imputing Confidential Characteristics and File Links in Longitudinal Linked Data. In Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra (eds). New York: Springer-Verlag, 290–297.

An, D. and Little, R.J.A. (2007). Multiple Imputation: an Alternative to Top Coding for Statistical Disclosure Control. Journal of the Royal Statistical Society, Series A, 170, 923–940.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., Zwick, T. (2007). A New Approach for Disclosure Control in the IAB Establishment Panel–Multiple Imputation for a Better Data Access. Technical Report, IAB Discussion Paper. No11/2007.

Graham, P. and Penny, R. (2005). Multiply Imputed Synthetic Data Files. Technical Report, University of Otago, http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm

Kennickell, A.B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In Record Linkage Techniques, W. Alvey and B. Jamerson (eds). Washington, D.C: National Academy Press, 248–267.

Kinney, S.K. (2007). Model Selection and Multivariate Inference Using Data Multiply Imputed for Disclosure Limitation and Nonresponse. Ph.D. thesis, Duke University, Department of Statistical Science.

Li, K.H., Meng, X.L., Raghunathan, T.E., 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.

Little, R.J.A., Liu, F., and Raghunathan, T.E. (2004). Statistical Disclosure Techniques Based on Multiple Imputation. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Persepectives, A. Gelman and X.L. Meng (eds). New York: John Wiley & Sons, 141–152.

Meng, X.L. and Rubin, D.B. (1992). Performing Likelihood Ratio Tests with Multiply-imputed Data Sets. Biometrika, 79, 103–111.

Mitra, R. and Reiter, J.P. (2006). Adjusting Survey Weights When Altering Identifying Design Variables via Synthetic Data. In Privacy in Statistical Databases 2006 (Lecture Notes in Computer Science), J. Domingo-Ferrar (ed.). New York: Springer-Verlag, 177–188.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models. Survey Methodology, 27, 85–96.

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. (2005a). Significance Tests for Multi-component Estimands from Multiply-imputed, Synthetic Microdata. Journal of Statistical Planning and Inference, 131, 365–377.

Reiter, J.P. (2005b). Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, 21, 441–462.

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 Mitra, R. (2008). Estimating Risks of Identification Disclosure in Partially Synthetic Data. Journal of Privacy and Confidentiality, 1.1, 99–110.

Reiter, J.P. and Raghunathan, T.E. (2007). The Multiple Adaptations of Multiple Imputation. Journal of the American Statistical Association, 102, 1462–1471.

Rubin, D.B. (1981). The Bayesian Bootstrap. The Annals of Statistics, 9, 130–134.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 462–468.

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

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

Woodcock, S.D. and Benedetto, G. (2006). Distribution-preserving Statistical Disclosure Limitation. Technical Report, Department of Economics, Simon Fraser University.

Received February 2008
Revised June 2009