

# Significance testing when using multiple imputation for missing data and disclosure limitation

BY SATKARTAR K. KINNEY AND JEROME P. REITER

*Institute of Statistics and Decision Sciences, Duke University*

*Box 90251, Durham, NC 27708, USA*

saki@stat.duke.edu   jerry@stat.duke.edu

## SUMMARY

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. In

this article, we propose significance tests for multicomponent hypotheses with such multiply-imputed datasets. The performance of these tests is illustrated with simulation studies.

*Some key words:* Confidentiality; Disclosure; Multiple Imputation; Nonresponse; Significance tests; Synthetic Data

## 1. INTRODUCTION

Many national statistical agencies, survey organizations, and researchers—henceforth all called agencies—disseminate microdata, i.e. data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents’ identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases and improvements in record linkage technologies have increased the risk of disclosure to the point where most agencies alter microdata before release (Reiter, 2004a). For example, agencies globally recode variables, such as releasing ages in five year intervals or top-coding incomes above 100,000 as “100,000 or more” (Willenborg & de Waal, 2001); they swap data values for randomly selected units (Dalenius & Reiss, 1982); or, they add random noise to continuous data values (Fuller, 1993). These strategies can reduce the utility of the released data, making some analyses impossible and severely distorting the results of others. They also complicate analyses for users. To analyze perturbed data properly, users should apply

the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach to disseminating public use data was suggested by Rubin (1993): release multiply-imputed, synthetic datasets. Specifically, he proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these datasets to the public. These are called fully synthetic datasets. Releasing fully synthetic data can protect confidentiality, since identification of units and their sensitive data is nearly impossible when the values in the released data are not actual, collected values. Furthermore, with appropriate synthetic data generation and the inferential methods developed by Raghunathan et al. (2003), users can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg et al. (1998), Raghunathan et al. (2003), Abowd & Lane (2004), and Reiter (2002, 2005a).

While no agencies have released fully synthetic datasets as of this writing, some have adopted a variant of the multiple imputation approach, suggested by Little (1993): release datasets comprising the units originally surveyed with some collected 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. For example, the U.S. Federal

Reserve Board protects data in the U.S. Survey of Consumer Finances by replacing monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). The U.S. Bureau of the Census and Abowd & Woodcock (2001, 2004) protect data in longitudinal, linked datasets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. Liu & Little (2002) and Little et al. (2004) present a general algorithm, named SMiKE, for simulating multiple values of key identifiers for selected units. These partially synthetic approaches are appealing because they promise to maintain the primary benefits of fully synthetic data—protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software—with decreased sensitivity to the specification of imputation models. Valid inferences from partially synthetic datasets can be obtained using the methods developed by Reiter (2003), whose rules for combining point and variance estimates differ from those of Rubin (1987) and also from those of Raghunathan et al. (2003). Other illustrations of partially synthetic data include Reiter (2005c) and Mitra & Reiter (2006).

When confidential datasets contain missing values, it is natural to use multiple imputation simultaneously for missing data and disclosure limitation. Reiter (2004b) describes a two-stage approach for this. First, the agency uses multiple imputation to fill in the missing data, generating  $m$  multiply-imputed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with  $n$  multiple imputations, ultimately releasing  $m * n$  multiply-imputed datasets. This approach is being used to create synthetic public use

files for the U.S. Survey of Income and Program Participation (Abowd et al., 2006). Generating the imputations in two stages enables users to estimate all sources of uncertainty – the sampling variability, the variability due to imputing missing data, and the variability due to replacing sensitive values. The rules of Rubin (1987) and Reiter (2003) do not apply in this two-stage imputation scheme. Appropriate rules for scalar estimands, similar in nature to those for nested multiple imputation for missing data (Shen, 2000; Harel & Schafer, 2003; Rubin, 2003), are presented in Reiter (2004b).

Often users of multiply-imputed data seek to test multicomponent null hypotheses, for example if several regression coefficients equal zero. Methods for performing such significance tests exist when multiple imputation is used for missing data only (Rubin, 1987; Li et al., 1991a,b; Meng & Rubin, 1992; Shen, 2000; Reiter, 2007) and for synthetic data only (Reiter, 2005b). In this paper, we propose such significance tests when multiple imputation is used to handle missing data and disclosure limitation simultaneously. The paper is organized as follows. Section 2 reviews the two-stage procedure of Reiter (2004b) and extends the appropriate combining rules to multivariate estimands. Section 3 describes a Wald test and a log-likelihood ratio test for testing multicomponent null hypotheses. Section 4 illustrates the properties of the Wald test using simulation studies. Section 5 provides 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  $l = 1, \dots, N$ . Let  $I = (I_1, \dots, I_N)$ , and let the sample size  $s = \sum I_l$ . 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  $s \times p$  sub-matrix 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  $D_o = (X, Y_{obs}, I, R)$ .

To generate the synthetic data, the agency first fills in values for  $Y_{mis}$  with draws from the conditional distribution of  $(Y_{mis} \mid D_o)$ , or approximations of that distribution such as those of Raghunathan et al. (2001). These draws are repeated independently  $i = 1, \dots, m$  times to obtain  $m$  completed datasets,  $D_c = \{D_c^{(i)} = (D_o, Y_{mis}^{(i)}), i = 1, \dots, m\}$ . Having dealt with the missing data, the agency limits disclosure risks by replacing selected values in each  $D_c^{(i)}$  with multiple imputations. For each  $D_c^{(i)}$ , imputations are made independently  $j = 1, \dots, n$  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  $Z = (Z_1, \dots, Z_s)$ . Let  $Y_{rep}^{(i,j)}$  be all the imputed (replaced) values in the  $j$ th synthetic dataset associated with  $D_c^{(i)}$ , and let  $Y_{nrep}^{(i)}$  be

all unchanged (unreplaced) values of  $D_c^{(i)}$ . The  $Y_{rep}^{(i,j)}$  are generated from the conditional distribution of  $(Y_{rep}^{(i,j)} \mid D_c^{(i)}, Z)$ , or a close approximation of it. Each synthetic dataset,  $D_s^{(i,j)}$ , then comprises  $(X, Y_{rep}^{(i,j)}, Y_{nrep}^{(i)}, I, R, Z)$ . The entire collection of  $M = mn$  datasets,  $D_s = \{D_s^{(i,j)}, i = 1, \dots, m; j = 1, \dots, n\}$ , with labels indicating the nests, is released to the public.

Reiter (2004b) derived an approximate  $t$ -distribution for inferences for scalar estimands from data multiply imputed in this manner. These results extend to multivariate estimands as follows. Let  $Q$  be a multivariate estimand, such as a vector of population means or regression coefficients. Let  $Q^{(i,j)}$  be the estimate of  $Q$  in data set  $D_s^{(i,j)}$ , and let  $U^{(i,j)}$  be the estimate of the variance associated with  $Q^{(i,j)}$ . The following quantities are needed for inferences.

$$\begin{aligned}\bar{Q} &= \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n Q^{(i,j)} = \frac{1}{m} \sum_{i=1}^m \bar{Q}^{(i)} \\ \bar{U} &= \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n U^{(i,j)} \\ \bar{W} &= \frac{1}{m} \sum_{i=1}^m \frac{1}{n-1} \sum_{j=1}^n (Q^{(i,j)} - \bar{Q}^{(i)})(Q^{(i,j)} - \bar{Q}^{(i)})' = \frac{1}{m} \sum_{i=1}^m W^{(i)} \\ B &= \frac{1}{m-1} \sum_{i=1}^m (\bar{Q}^{(i)} - \bar{Q})(\bar{Q}^{(i)} - \bar{Q})'.\end{aligned}$$

Finally, let  $B_\infty = \lim B$  as  $m \rightarrow \infty$  and  $n \rightarrow \infty$ , and let  $W_\infty^{(i)} = \lim W^{(i)}$  as  $n \rightarrow \infty$ . Let  $\bar{W}_\infty = \sum_{i=1}^m W_\infty^{(i)} / m$ .

Assuming the conditions for valid inferences under multiple imputation for missing data (Rubin, 1987), we have

$$(Q \mid D_c, B_\infty, \bar{W}_\infty) \sim N(\bar{Q}_c, \bar{U} + (1 + 1/m)B_\infty) \quad (1)$$

where  $\bar{Q}_c = \sum_{i=1}^m Q_c^{(i)} / m$  and each  $Q_c^{(i)}$  is the estimate of  $Q$  that would be obtained from its corresponding  $D_c^{(i)}$  prior to replacement of confidential values. An implicit assumption here is that the  $U^{(i,j)}$  have sufficiently low variability so that  $U^{(i,j)} \simeq U_c^{(i)}$ , where  $U_c^{(i)}$  is the variance estimate of  $Q_c^{(i)}$  computed from  $D_c^{(i)}$ . Similarly, we assume that the  $U_c^{(i)} \simeq U$ , and hence  $\bar{U} \simeq U$ , 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 the usual assumptions in multiple imputation, motivated by the fact that posterior variances generally have lower order variability than posterior means (Rubin, 1987, p.89).

Assuming the conditions for valid inferences under partially synthetic data (Reiter, 2003), we have

$$(Q_c^{(i)} | D_s, B_\infty, W_\infty^{(i)}) \sim N(\bar{Q}^{(i)}, W_\infty^{(i)} / n). \quad (2)$$

Integrating (1) and (2) with respect to the  $Q_c^{(i)}$ , we have

$$(Q | D_s, B_\infty, \bar{W}_\infty) \sim N(\bar{Q}, T_\infty) \quad (3)$$

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 the  $B_\infty$  and the  $W_\infty^{(i)}$  are not known and must be integrated out of (3). Extending the theory for scalar quantities in Reiter (2004b), the sampling distribution of each  $\bar{Q}^{(i)}$  is

$$(\bar{Q}^{(i)} | D_o, B_\infty, \bar{W}_\infty) \sim N(Q_o, B_\infty + \bar{W}_\infty / n) \quad (4)$$

where  $Q_o$  is the estimate of  $Q$  that would be obtained from  $D_o$ . Hence, assuming that  $W_\infty^{(i)} = \bar{W}_\infty$  for all  $i$  and diffuse prior distributions on all parameters, from (4) and (2), we have

$$\{B(B_\infty + \bar{W}_\infty/n)^{-1}|D_s, \bar{W}_\infty\} \sim Wi(m-1, I) \quad (5)$$

$$\{W^{(i)}(W_\infty^{(i)})^{-1}|D_s\} \sim Wi(n-1, I). \quad (6)$$

For sufficiently large  $s$ ,  $m$ , and  $n$ , we can replace  $B_\infty$  and  $\bar{W}_\infty$  with their approximate expected values, resulting in the variance estimate  $T = (1 + 1/m)B - (1/n)\bar{W} + \bar{U}$ ; thus, inferences for  $Q$  can be based on  $(\bar{Q} - Q) \sim N(0, T)$ . The fractional increase in variance due to missing data can be computed from  $D_s$  as  $(1 + 1/m)(B - \bar{W}/n)\bar{U}^{-1}$  and the fractional increase due to replacement data as  $\{\bar{W}/(mn)\}\bar{U}^{-1}$ . For each of these, the average fractional increase across components of  $Q$  can be obtained by averaging the diagonal elements of these matrices.

### 3. SIGNIFICANCE TESTING

Using the  $M$  released datasets, an analyst seeks to test the null hypothesis  $Q = Q_0$  for some  $k$ -component estimand  $Q$ , for example to test if  $k$  regression coefficients equal zero. Given the normal approximation for inferences about  $Q$ , it may appear reasonable to use a Wald test with test statistic  $(Q - Q_0)'T^{-1}(Q - Q_0)$ ; 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. This is a problem for small  $m$

even when no synthetic data are generated (Rubin, 1987; Li et al., 1991a,b). The instability in  $T$  can be avoided by making  $m$  and  $n$  large; however, that is typically impractical.

We propose two approaches to significance testing for multivariate  $Q$ . The first is a test based on Wald statistics. This requires access to all elements of the  $U^{(i,j)}$  matrices. The second is a test based on likelihood ratio statistics, which is most useful when the elements of the  $U^{(i,j)}$  are not available, or when the dimension of  $U^{(i,j)}$  makes working with Wald statistics too cumbersome. For both tests, we first present the test statistic and its reference distribution, followed by the derivation.

### 3.1. Wald test

The key idea in the derivation of the Wald test statistic is to reduce the number of unknown parameters in  $B_\infty$  and the  $W_\infty^{(i)}$  by assuming (i) equal fractions of missing information on each component of  $Q$ , and (ii) equal fractions of replaced information on each component of  $Q$ . Equivalently, the  $B_\infty$  and  $\bar{W}_\infty$  are proportional to  $\bar{U}$ . Similar proportionality assumptions are used in multiple imputation for missing data only (Rubin, 1987; Li et al., 1991a,b; Shen, 2000) and for synthetic data only (Reiter, 2005b). We derive a reference  $F$ -distribution for the test statistic following the moment-matching approach proposed by Li et al. (1991b).

The statistic for the Wald test is

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

where

$$r^{(b)} = (1 + 1/m)\text{tr}(B\bar{U}^{-1})/k \quad (7)$$

$$r^{(w)} = (1/n)\text{tr}(\bar{W}\bar{U}^{-1})/k. \quad (8)$$

The reference distribution is approximated by an  $F$ -distribution with  $k$  degrees of freedom in the numerator and  $w_s$  degrees of freedom in the denominator, where

$$w_s = 4 + \left\{ 1 + \frac{r^{(b)}\nu_b}{\nu_b - 2} - \frac{r^{(w)}\nu_w}{\nu_w - 2} \right\}^2 / \left\{ \frac{(r^{(b)}\nu_b)^2}{(\nu_b - 2)^2(\nu_b - 4)} + \frac{(r^{(w)}\nu_w)^2}{(\nu_w - 2)^2(\nu_w - 4)} \right\} \quad (9)$$

for  $\nu_b > 4$  and  $\nu_w > 4$ , and  $\nu_b = k(m - 1)$  and  $\nu_w = km(n - 1)$ . The approximate p-value for testing  $Q = Q_0$  is given by  $pr(F_{k,w_s} > S)$ . When  $n$  is large, or when  $\bar{W}$  is small,  $S$  and  $w_s$  approximately equal the test statistic and degrees of freedom developed by Li et al. (1991b) for multiple imputation for missing data only.

When  $\nu_b \leq 4$  or  $\nu_w \leq 4$ ,  $w_s$  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 of  $T < 0$  (Reiter, 2006). In such cases, we suggest an alternate denominator degrees of freedom,

$$w_s^* = \left\{ \frac{(r^{(b)})^2}{\nu_b(1 + r^{(b)} - r^{(w)})^2} + \frac{(r^{(w)})^2}{\nu_w(1 + r^{(b)} - r^{(w)})^2} \right\}^{-1}, \quad (10)$$

which is a generalization of the degrees of freedom used in the  $t$ -distribution of Reiter (2004b) for inferences for scalar  $Q$ .

We next present the derivation of  $S$  and its reference  $F$ -distribution. Conditional on  $T_\infty$  and using (3), the p-value for testing  $Q = Q_0$  is  $pr(\chi_k^2 > (Q_0 - \bar{Q})'T_\infty^{-1}(Q_0 - \bar{Q}))$ , where  $\chi_k^2$  is a chi-squared random variable on  $k$  degrees of freedom. Since  $T_\infty$  is generally

not known, we obtain the p-value by averaging over the distributions of  $(B_\infty|D_s, \bar{W}_\infty)$  and  $(\bar{W}_\infty|D_s)$  in (5) and (6), resulting in

$$\int pr\{\chi_k^2 > (Q_0 - \bar{Q})'T_\infty^{-1}(Q_0 - \bar{Q})|D_s, B_\infty, \bar{W}_\infty\}pr(B_\infty|D_s, \bar{W}_\infty)pr(\bar{W}_\infty|D_s)dB_\infty d\bar{W}_\infty.$$

This integral can be evaluated numerically, but it is desirable to have a closed-form approximation. We assume that  $B_\infty = r_\infty^{(b)}\bar{U}_\infty$  and  $W_\infty^{(i)} = r_\infty^{(w)}\bar{U}_\infty$  for all  $i$ , where  $r_\infty^{(w)}$  and  $r_\infty^{(b)}$  are scalar quantities. Assuming  $\bar{U}_\infty = \bar{U}$ , and that  $W_\infty^{(i)} = \bar{W}_\infty$  for all  $i$ , we have  $B_\infty = r_\infty^{(b)}\bar{U}$  and  $\bar{W}_\infty = r_\infty^{(w)}\bar{U}$ . Using those assumptions, the p-value is

$$\begin{aligned} & \int pr\left\{\chi_k^2 > \frac{(Q_0 - \bar{Q})'\bar{U}^{-1}(Q_0 - \bar{Q})}{1 + (1 + 1/m)r_\infty^{(b)} + r_\infty^{(w)}/(mn)}|D_s, r_\infty^{(b)}, r_\infty^{(w)}\right\} \times \\ & \quad pr(r_\infty^{(b)}|D_s, r_\infty^{(w)})pr(r_\infty^{(w)}|D_s)dr_\infty^{(b)}dr_\infty^{(w)} \\ & = \int pr\left\{(\chi_k^2/k)\frac{1 + (1 + 1/m)r_\infty^{(b)} + r_\infty^{(w)}/(mn)}{(1 + r_\infty^{(b)} - r_\infty^{(w)})} > S|D_s, r_\infty^{(b)}, r_\infty^{(w)}\right\} \times \\ & \quad pr(r_\infty^{(b)}|D_s, r_\infty^{(w)})pr(r_\infty^{(w)}|D_s)dr_\infty^{(b)}dr_\infty^{(w)}. \end{aligned} \quad (11)$$

The conditional distributions of  $r_\infty^{(b)}$  and  $r_\infty^{(w)}$  can be obtained from (2), (4), (5), and (6).

Applying multivariate normal theory, we obtain:

$$\left\{\frac{k(m-1)\text{tr}(B\bar{U}^{-1})/k}{r_\infty^{(b)} + r_\infty^{(w)}/n}|D_s, r_\infty^{(w)}\right\} \sim \chi_{k(m-1)}^2 \quad (12)$$

$$\left\{\frac{km(n-1)\text{tr}(\bar{W}\bar{U}^{-1})/k}{r_\infty^{(w)}}|D_s\right\} \sim \chi_{km(n-1)}^2. \quad (13)$$

Substituting (12) and (13) into (11), after some algebra we obtain

$$pr\left\{(\chi_k^2/k)\frac{1 + v_b r^{(b)}/\chi_{v_b}^2 - v_w r^{(w)}/\chi_{v_w}^2}{1 + r^{(b)} - r^{(w)}} > S\right\}. \quad (14)$$

We approximate the random variable in (14) as proportional to a  $F$ -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 (14). Equivalently, we approximate  $(1 + \chi_{\nu_b}^{-2} \nu_b r^{(b)} - \chi_{\nu_w}^{-2} \nu_w r^{(w)})$  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  $\eta \chi_{w_s}^{-2}$ . Using iterated expectations and variances, we have

$$E(\eta \chi_{w_s}^{-2}) = \frac{\eta}{w_s - 2} \simeq 1 + \frac{\nu_b r^{(b)}}{\nu_b - 2} - \frac{\nu_w r^{(w)}}{\nu_w - 2}$$

and

$$\begin{aligned} E\{(\eta \chi_{w_s}^{-2})^2\} &= \frac{\eta^2}{(w_s - 2)(w_s - 4)} \\ &\simeq \frac{2(\nu_w r^{(w)})^2}{(\nu_w - 2)^2(\nu_w - 4)} + \frac{2(\nu_b r^{(b)})^2}{(\nu_b - 2)^2(\nu_b - 4)} + \left(1 + \frac{\nu_b r^{(b)}}{\nu_b - 2} - \frac{\nu_w r^{(w)}}{\nu_w - 2}\right)^2. \end{aligned}$$

Solving yields the expression in (9) for  $w_s$  and  $\delta = \{(w_s - 2)/w_s\} \{1 + \nu_b r^{(b)}/(\nu_b - 2) - \nu_w r^{(w)}/(\nu_w - 2)\} / (1 + r^{(b)} - r^{(w)})$ . When  $\nu_b$  and  $\nu_w$  are sufficiently large,  $\delta \simeq 1$ , and the approximate p-value is  $pr(F_{k,w_s} > S)$ .

### 3.2. Log-likelihood ratio test

Meng & Rubin (1992) developed an alternative test for conventional multiple imputation for missing data, based on the set of log-likelihood ratio test statistics from the completed datasets. This was extended to nested multiple imputation for missing data only by Shen (2000) and to synthetic data only by Reiter (2005b). In this section, we extend this test to the case of missing and synthetic data handled simultaneously.

Following the notation in Schafer (1997), let  $\psi$  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

$D_s^{(i,j)}$  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  $D_s^{(i,j)}$  as  $d'(a, b|D_s^{(i,j)}) = 2 \log f(D_s^{(i,j)}|a) - 2 \log f(D_s^{(i,j)}|b)$ .

The test statistic is

$$\tilde{S} = \bar{L}/\{k(1 + \tilde{r}^{(b)} - \tilde{r}^{(w)})\} \quad (15)$$

where  $\tilde{r}^{(b)} = \{(m+1)(\bar{L}_m - \bar{L})\}/\{k(m-1)\}$ ,  $\tilde{r}^{(w)} = (\bar{l} - \bar{L}_m)/\{k(n-1)\}$ ,  $\bar{L} = \sum_{i=1}^m \sum_{j=1}^n d'(\bar{\psi}_0, \bar{\psi}|D_s^{(i,j)})/(mn)$ ,  $\bar{L}_m = \sum_{i=1}^m \sum_{j=1}^n d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)}|D_s^{(i,j)})/(mn)$ , and  $\bar{l} = \sum_{i=1}^m \sum_{j=1}^n d'(\hat{\psi}_0^{(i,j)}, \hat{\psi}^{(i,j)}|D_s^{(i,j)})/(mn)$ .

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 (9) with the terms  $r^{(b)}$  and  $r^{(w)}$  replaced by  $\tilde{r}^{(b)}$  and  $\tilde{r}^{(w)}$ . When  $\nu_b \leq 4$  or  $\nu_w \leq 4$ , we use the denominator degrees of freedom in (10), substituting in  $\tilde{r}^{(b)}$  and  $\tilde{r}^{(w)}$  as above.

The derivation parallels the strategy of Meng & Rubin (1992), namely (i) find a statistic asymptotically equivalent to  $S$  based only 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  $d'(\hat{\psi}_0^{(i,j)}, \hat{\psi}^{(i,j)}|D_s^{(i,j)})$ . Fur-

thermore, 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

$$S^* = \frac{\frac{\bar{d}}{k} - (n-1)r^{(w)} - (m-1)r^{(b)}/(m+1)}{1 + r^{(b)} - r^{(w)}} \quad (16)$$

where  $r^{(b)}$  and  $r^{(w)}$  are defined in (7) and (8). 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,

$$\begin{aligned} \bar{d} &= \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (Q^{(i,j)} - \bar{Q}^{(i)})'(Q^{(i,j)} - \bar{Q}^{(i)}) + \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\bar{Q}^{(i)} - \bar{Q})'(\bar{Q}^{(i)} - \bar{Q}) + \bar{Q}'\bar{Q} \\ &= k(n-1)r^{(w)} + \frac{k(m-1)}{m+1}r^{(b)} + \bar{Q}'\bar{Q}. \end{aligned}$$

Substituting the above expression into (16) 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

$$\begin{aligned} r^{(b)} &= \frac{(m+1)}{km(m-1)} \sum_{i=1}^m (\bar{Q}^{(i)} - \bar{Q})'(\bar{Q}^{(i)} - \bar{Q}) = \frac{(m+1)}{km(m-1)} \left\{ \sum_{i=1}^m (\bar{Q}^{(i)'}\bar{Q}^{(i)}) - m\bar{Q}'\bar{Q} \right\} \\ &\simeq \frac{(m+1)}{k(m-1)} \left( \sum_{i=1}^m \bar{d}^{(i)}/m - \hat{d} \right) = r_w^{(b)} \end{aligned}$$

since  $\sum_{i=1}^m \bar{d}^{(i)}/m$  is asymptotically equivalent to  $\sum_{i=1}^m (\bar{Q}^{(i)'}\bar{Q}^{(i)})$ , and  $\hat{d}$  is asymptotically

equivalent to  $\bar{Q}'\bar{Q}$ . For  $r^{(w)}$ , we have

$$\begin{aligned}
r^{(w)} &= \frac{1}{kmn(n-1)} \sum_{i=1}^m \sum_{j=1}^n (Q^{(i,j)} - \bar{Q}^{(i)})' (Q^{(i,j)} - \bar{Q}^{(i)}) \\
&= \frac{1}{kmn(n-1)} \left\{ \sum_{i=1}^m \sum_{j=1}^n (Q^{(i,j)'} Q^{(i,j)}) - n \sum_{i=1}^m (\bar{Q}^{(i)'} \bar{Q}^{(i)}) \right\} \\
&\simeq \frac{1}{k(n-1)} \left( \bar{d} - \sum_{i=1}^m \bar{d}^{(i)}/m \right) = r_w^{(w)}.
\end{aligned}$$

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 (15) 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  $d'(\hat{\psi}^{(i,j)}, \hat{\psi}_0^{(i,j)} | D_s^{(i,j)})$ . 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 & Rubin (1992) and Shen (2000), for quadratic complete-data log-likelihood functions, we have

$$\begin{aligned}
d'(\bar{\psi}_0, \bar{\psi} | D_s^{(i,j)}) &\simeq d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}, U^{(i,j)}) \\
d'(\bar{\psi}_0^{(i)}, \bar{\psi}^{(i)} | D_s^{(i,j)}) &\simeq d(Q^{(i,j)}, U^{(i,j)}) - d(Q^{(i,j)} - \bar{Q}^{(i)}, U^{(i,j)}).
\end{aligned}$$

Thus, we have

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

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.

#### 4. SIMULATION STUDIES

In this section, we evaluate the performance of the Wald test for multicomponent estimands using simulations. Since the likelihood ratio test is asymptotically equivalent to the Wald test, for large samples it should have similar performance.

For sample size  $s = 1000$ , we simulate the complete data  $\{Y_0, Y_1, \dots, Y_{20}\}$  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  $\{Y_1, \dots, Y_{20}\}$  missing completely at random and  $Y_0$  always fully observed. We obtain the set of completed datasets,  $D_c$ , by drawing values of the missing data from  $f(Y_1, \dots, Y_{20}|D_o)$ , 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  $D_s^{(i,j)}$  are drawn independently from  $f(Y_0|D_c^{(i)})$ . We vary the number of imputations according to  $m \in (4, 8)$  and  $r \in (2, 4, 8)$ . By design, this simulation satisfies both proportionality assumptions.

The hypothesis that we test is  $H_0 : Q = 0$ , where  $Q$  is the vector of coefficients for the regression of  $Y_0$  on  $Y_1, \dots, Y_k$ , excluding the intercept, for  $k \in (5, 10, 20)$ . As this null hypothesis is true, we expect that for a given significance level  $\alpha$  that  $H_0$  will be rejected

100 $\alpha$ % of the time. Table 1 summarizes the simulated nominal significance levels of the Wald test using 10000 runs of the simulation for each combination of  $m$ ,  $n$ , and  $k$ , for  $\alpha \in (.01, .05, .10)$ . The simulated significance levels are close the desired significance levels. The rates are low when  $r = 2$ , suggesting the tests may be conservative in these cases. The conventional Wald test using covariance matrix  $T = (1+1/m)B - (1/n)\bar{W} + \bar{U}$  requires a much larger number of imputations to yield correct levels. As shown in Table 2, this test has dramatically high rejection rates for the more realistic values of  $m$  and  $n$  used.

Li et al. (1991a) show for multiple imputation for missing data only, that similar Wald tests based on the proportionality assumption are robust in cases of practical interest even when the proportionality assumption fails. To evaluate the robustness of the test to violations of the proportionality assumption in the synthetic data, we next perform a simulation in which the proportionality assumption is not met for the synthetic replacement data. In this second simulation,  $Y_0, \dots, Y_{10}$  are replaced in entirety, and  $Y_{11}, \dots, Y_{20}$  are left intact. The imputations are generated from  $D_c$  by taking draws from  $(Y_{10}|Y_{11}, \dots, Y_{20})$ ,  $(Y_9|Y_{10}, \dots, Y_{20})$ ,  $\dots$ ,  $(Y_0|Y_1, \dots, Y_{20})$ . We set  $k = 20$  and test  $H_0 : Q = 0$ , where  $Q$  is the vector of coefficients from the regression of  $Y_{20}$  on  $Y_0, \dots, Y_{19}$ . Table 3 gives the nominal rejection rates for this scenario, which are seen to be similar to those in Table 1.

## 5. CONCLUDING REMARKS

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. We recommend that analysts use the Wald test when possible, because the likelihood ratio test involves further approximations. Software distributed with partially synthetic datasets can make the Wald test the default option.

The simulations suggest that the Wald tests have 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 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.

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 paper 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.

#### ACKNOWLEDGEMENT

This work was supported by a grant from the U.S. National Science Foundation.

#### REFERENCES

- ABOWD, J. M. & LANE, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer & V. Torra, eds., *Privacy in Statistical Databases*. New York: Springer-Verlag, 282–289.
- ABOWD, J. M., STINSON, M. H. & BENEDETTO, G. L. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program.
- ABOWD, J. M. & WOODCOCK, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz & J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North-Holland, 215–277.
- ABOWD, J. M. & WOODCOCK, S. D. (2004). Multiply-imputing confidential character-

- istics and file links in longitudinal linked data. In J. Domingo-Ferrer & V. Torra, eds., *Privacy in Statistical Databases*. New York: Springer-Verlag, 290–297.
- DALENIUS, T. & REISS, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6** 73–85.
- FIENBERG, S. E., MAKOV, U. E. & STEELE, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14** 485–502.
- FULLER, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9** 383–406.
- HAREL, O. & SCHAFER, J. (2003). Multiple imputation in two stages. In *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*.
- KENNICHELL, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey & B. Jamerson, eds., *Record Linkage Techniques, 1997*. Washington, D.C.: National Academy Press, 248–267.
- LI, K. H., RAGHUNATHAN, T. E., MENG, X. L. & 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. & 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., LIU, F. & RAGHUNATHAN, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman & X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. New York: John Wiley & Sons, 141–152.
- LITTLE, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9** 407–426.
- LIU, F. & LITTLE, R. J. A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*. 2133–2138.
- MENG, X. L. & RUBIN, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79** 103–111.
- MITRA, R. & REITER, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrar, ed., *Privacy in Statistical Databases 2006 (Lecture Notes in Computer Science)*. New York: Springer-Verlag, 177–188.
- RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J. & SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27** 85–96.
- RAGHUNATHAN, T. E., REITER, J. P. & RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19** 1–16.

- REITER, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18** 531–544.
- REITER, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29** 181–189.
- REITER, J. P. (2004a). New approaches to data dissemination: A glimpse into the future (?). *Chance* **17** 12–16.
- REITER, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30** 235–242.
- REITER, J. P. (2005a). 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. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131** 365–377.
- REITER, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21** 441–462.
- REITER, J. P. (2006). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. Tech. rep., ISDS, Duke University.

- REITER, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika (forthcoming)* .
- 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. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57** 3–18.
- SCHAFFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- SHEN, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics.
- WILLENBORG, L. & DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

		$\alpha = .01$			$\alpha = .05$			$\alpha = .10$		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$m = 4$	$n = 2$	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

Table 1: Nominal rejection rates for given significance level  $\alpha$  using  $pr(F_{k,w_s} > S)$  when proportionality assumptions are satisfied

		$\alpha = .01$			$\alpha = .05$			$\alpha = .10$		
		$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
$m = 4$	$n = 2$	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

Table 2: Nominal rejection rates for given significance level  $\alpha$  using Wald test based on covariance matrix  $T$  when proportionality assumptions are satisfied

		$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
	$n = 2$	1.1	5.3	10.4
$m = 4$	$n = 4$	1.2	5.8	10.6
	$n = 8$	1.5	5.3	10.4
	$n = 2$	1.2	5.7	10.9
$m = 8$	$n = 4$	1.3	5.5	10.5

Table 3: *Nominal rejection rates for  $k = 20$  and given significance level  $\alpha$  using  $pr(F_{k,w_s} > S)$  when the synthetic data proportionality assumption is not satisfied*