Biometrika (1999), **86,** 4, *pp.* 948–955
© 1999 Biometrika Trust
Printed in Great Britain

Small-sample degrees of freedom with multiple imputation

By JOHN BARNARD and DONALD B. RUBIN

Department of Statistics, Harvard University, 1 Oxford Street, Cambridge,
Massachusetts 02138, U.S.A.

barnard@stat.harvard.edu rubin@stat.harvard.edu

Summary

An appealing feature of multiple imputation is the simplicity of the rules for combining the multiple complete-data inferences into a final inference, the repeated-imputation inference (Rubin, 1987). This inference is based on a t distribution and is derived from a Bayesian paradigm under the assumption that the complete-data degrees of freedom, ν_{com}, are infinite, but the number of imputations, m, is finite. When ν_{com} is small and there is only a modest proportion of missing data, the calculated repeated-imputation degrees of freedom, ν_m, for the t reference distribution can be much larger than ν_{com}, which is clearly inappropriate. Following the Bayesian paradigm, we derive an adjusted degrees of freedom, $\tilde{\nu}_m$, with the following three properties: for fixed m and estimated fraction of missing information, $\tilde{\nu}_m$ monotonically increases in ν_{com}; $\tilde{\nu}_m$ is always less than or equal to ν_{com}; and $\tilde{\nu}_m$ equals ν_m when ν_{com} is infinite. A small simulation study demonstrates the superior frequentist performance when using $\tilde{\nu}_m$ rather than ν_m.

Some key words: Bayesian inference; Fraction of missing information; Missing at random; Missing data mechanism; Repeated imputation.

1. Introduction

Multiple imputation (Rubin, 1987) is becoming a standard approach for handling missing data problems because of the availability of software, e.g. Schafer (1997) and Statistical Solutions (1997), its successful adoption for large-scale examples (Schafer, Khare & Ezzati-Rice, 1993), and theoretical and applied evidence of its reliability in producing valid randomisation-based inferences using the standard rules for combining the inferences from each set of imputations; see Meng & Rubin (1992) and references in Rubin (1996).

These basic rules, the ‘repeated imputation inferences’, are derived in Rubin & Schenker (1986) and Rubin (1987, Ch. 3) based on a Bayesian analysis that assumes that imputations are independent draws from the posterior predictive distribution of the missing values, and that the size of the complete dataset is large, in the sense that, if there were no missing values, inferences would be based on large-sample methods; i.e. the degrees of freedom for standard errors and denominators of test statistics would effectively be set at infinity. Repeated imputation inferences are based on a t reference distribution with associated degrees of freedom (Rubin, 1987, eqn (3.1.6))

$$\nu_m = (m-1)\hat{\gamma}_m^{-2}, \tag{1}$$

where

$$\hat{\gamma}_m = (1+m^{-1})\,\text{tr}(B_m T_m^{-1})/k \tag{2}$$

is approximately the Bayesian fraction of missing information for the unknown quantity of interest (Rubin, 1987, p. 93), B_m and T_m, the between and total variances, are defined in § 2, k is the dimension of the estimand, and m is the number of imputations.

In small datasets, however, it can be unsatisfactory to set degrees of freedom to infinity, especially when there is little missing information for the estimand, because ν_m can then be many times the degrees of freedom available if there were no missing data. This situation indicates the need for a new expression for the multiple imputation degrees of freedom that does not rely on a large complete-data sample.

Here we provide a principled adjustment to (1) such that, for fixed m and $\hat{\gamma}_m$, the resulting degrees of freedom, $\tilde{\nu}_m$, monotonically increases in the complete-data degrees of freedom, ν_{com}, and is always less than ν_{com}. This adjusted degrees of freedom, which reduces to (1) when $\nu_{\text{com}} = \infty$, is

$$\tilde{\nu}_m = \left(\frac{1}{\nu_m} + \frac{1}{\hat{\nu}_{\text{obs}}}\right)^{-1} = \nu_m\left(1 + \frac{\nu_m}{\hat{\nu}_{\text{obs}}}\right)^{-1} = \nu_{\text{com}}\left[\{\lambda(\nu_{\text{com}})(1-\hat{\gamma}_m)\}^{-1} + \frac{\nu_{\text{com}}}{\nu_m}\right]^{-1}, \qquad (3)$$

where

$$\lambda(\nu) = \frac{\nu+1}{\nu+3} \qquad (4)$$

and

$$\hat{\nu}_{\text{obs}} = \lambda(\nu_{\text{com}})\nu_{\text{com}}(1-\hat{\gamma}_m) \qquad (5)$$

is the estimated observed-data degrees of freedom. The two latter expressions in (3) show that $\tilde{\nu}_m$ is always less than or equal to both ν_m, given by (1), and ν_{com}.

We begin by defining notation and reviewing the repeated-imputation inference procedure in § 2. In § 3 we derive the degrees of freedom when there is an infinite number of imputations and the missing data are assumed to be missing completely at random. In § 4 we build on the results in § 3, obtaining the expression for $\tilde{\nu}_m$ given in (3) when there is only a finite number of imputations. In § 5 we conclude with a brief simulation demonstrating the superior frequentist operating characteristics of procedures based on (3) rather than (1).

2. Repeated-imputation procedure

Let Y be the complete data, i.e. what we would observe in the absence of any missing data; let Y_{obs} represent the observed components of Y, and Y_{mis} the missing components, that is, $Y \equiv (Y_{\text{obs}}, Y_{\text{mis}})$. We assume that, with complete data, valid inference about a k-component quantity Q, possibly a model parameter or a finite population characteristic, would be based on the standard large-sample statement

$$(Q - \hat{Q}) \sim N(0, U), \qquad (6)$$

where $\hat{Q} \equiv \hat{Q}(Y)$ is an estimator of Q, and $U \equiv U(Y)$ is its associated variance. We allow this assumption to be relaxed by replacing the normal distribution in (6) with a t distribution, as is often appropriate subasymptotically.

The basic idea of multiple imputation is to fill in the missing data multiple times with values drawn from some distribution that predicts the missing values given the observed data and other available information. Each draw of Y_{mis} is an imputation. To obtain a repeated-imputation inference, which treats the m imputations, $Y_{\text{mis}}^{(1)}$ through $Y_{\text{mis}}^{(m)}$, as repeated independent draws of Y_{mis} from a Bayesian prediction model (Rubin, 1987, p. 75), an analyst of the multiply imputed dataset performs three steps. The first step is to calculate the complete-data statistics, $\hat{Q}_{*l} \equiv \hat{Q}(Y^{(l)})$ and $U_{*l} \equiv U(Y^{(l)})$, for each of the m completed datasets, $Y^{(l)} \equiv (Y_{\text{obs}}, Y_{\text{mis}}^{(l)})$, for $l = 1, \ldots, m$. In the second step, the analyst computes the estimate of Q,

$$\bar{Q}_m = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_{*l},$$

and its associated variance estimate

$$T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right)B_m, \qquad (7)$$

where

$$\bar{U}_m = \frac{1}{m} \sum_{l=1}^{m} U_{*l}$$

measures the within-imputation variability, and

$$B_m = \frac{1}{m-1} \sum_{l=1}^{m} (\hat{Q}_{*l} - \bar{Q}_m)(\hat{Q}_{*l} - \bar{Q}_m)^{\mathrm{T}}$$

measures the between-imputation variability; the adjustment $(1 + m^{-1})$ is due to the additional variance from using $\bar{Q}_m$ rather than $\bar{Q}_\infty$.

Finally, interval estimates and significance levels for Q are based on a k-component Student-t reference distribution on v_m degrees of freedom,

$$T_m^{-1/2}(Q - \bar{Q}_m) \sim t_{v_m}, \tag{8}$$

where v_m is given in (1). The resulting inference is called the repeated-imputation inference (Rubin, 1987, p. 75) because it is derived assuming the imputations are independent repetitions from a posterior predictive distribution, thereby distinguishing it from other multiple-imputation inferences, e.g. as discussed in Meng (1994).

3. RESULT FOR INFINITE m

Suppose first that an infinite number of multiple imputations were created. Thus $\bar{Q}_m$ is $\bar{Q}_\infty$, $\bar{U}_m$ is $\bar{U}_\infty$, and B_m is B_∞, as in Rubin (1987, § 3.2). Here we assume that, with complete data, valid inference about a k-component quantity Q would be based on the statement

$$(Q - \hat{Q}) \sim t_{v_{\mathrm{com}}}(0, U), \tag{9}$$

where $t_{v_{\mathrm{com}}}(0, U)$ is the t distribution with location 0, squared scale U and v_{com} degrees of freedom, i.e. we replace the large-sample normality assumption in (6) with a t assumption. In addition, we assume that the posterior distribution of Q, that is the distribution of Q given Y_{obs}, or equivalently given $\bar{Q}_\infty$, $\bar{U}_\infty$ and B_∞, is

$$(Q \mid \bar{Q}_\infty, B_\infty, \bar{U}_\infty) \sim t_{v_{\mathrm{obs}}}(\bar{Q}_\infty, T_\infty), \tag{10}$$

where v_{obs} is the observed data degrees of freedom and $T_\infty = \bar{U}_\infty + B_\infty$. Equation (10) can be viewed as a modification of (3.3.1) in Rubin (1987).

Since v_{obs} is generally unknown, we approximate it. Under (9) and (10) the Bayesian fraction of missing information (Rubin, 1987, p. 86) in an average sense is

$$\gamma_\infty = 1 - \mathrm{tr}[\{\lambda(v_{\mathrm{obs}})\bar{U}_\infty\}\{\lambda(v_{\mathrm{com}})T_\infty\}^{-1}]/k,$$

where $\lambda(v)$ is defined in (4). Since $\lambda(v_{\mathrm{com}}) < \lambda(v_{\mathrm{obs}})/\lambda(v_{\mathrm{com}})$ for small to moderate fractions of missing information, we replace $\lambda(v_{\mathrm{obs}})/\lambda(v_{\mathrm{com}})$ with $\lambda(v_{\mathrm{com}})$, yielding an approximation to the average fraction of missing information that is free of v_{obs}:

$$\tilde{\gamma}_\infty = 1 - \lambda(v_{\mathrm{com}})\, \mathrm{tr}(\bar{U}_\infty T_\infty^{-1})/k.$$

When the missing data process is missing completely at random (Rubin, 1976), the missing information simply reflects an effective reduction in sample size; hence, under this scenario, the observed data degrees of freedom, v_{obs}, are, in an average and typically conservative sense, $v_{\mathrm{com}}(1 - \tilde{\gamma}_\infty)$ or equivalently

$$v_{\mathrm{obs}} = v_{\mathrm{com}}^{*}\, \mathrm{tr}(\bar{U}_\infty T_\infty^{-1})/k, \tag{11}$$

where $v_{\mathrm{com}}^{*} = \lambda(v_{\mathrm{com}})v_{\mathrm{com}}$.

For example, consider a simple random sample $y_1, \ldots, y_n$ from a population with unknown mean μ, where $n - n_{\text{obs}}$ observations are missing completely at random, with $n \geqslant n_{\text{obs}} \geqslant 2$. For estimation of μ, we have $v_{\text{com}} = n - 1$ and the correct degrees of freedom are $n_{\text{obs}} - 1$; $\bar{U}_\infty / T_\infty = n_{\text{obs}}/n$, and thus $v_{\text{obs}} = n_{\text{obs}}\{(n-1)/(n+2)\}$, which for small amounts of missing information is close to $n_{\text{obs}} - 1$, and is always between n_{obs} and $n_{\text{obs}} - 3$.

4. Result for modest m

As in Rubin (1987, Ch. 3), we first assume that B_∞ is known and find the conditional posterior distribution of Q, having integrated over $\bar{Q}_\infty$ given the set of repeated-imputation statistics $S_m = \{\hat{Q}_{*l}, U_{*l};\ l = 1, \ldots, m\}$. The repeated draws of the statistics $(\hat{Q}_{*l}, U_{*l})$ based on the imputed data are independent and identically distributed draws from the joint posterior distribution of $(\hat{Q}, U)$, where, although the U_{*l} cannot be truly considered constant as when $v_{\text{obs}} = \infty$, they still tend to have relatively less variability than the $\hat{Q}_{*l}$; hence, we still treat $\bar{U}_m$ as fixed; i.e. we let $\bar{U}_m = \bar{U}_\infty$. Since the mean of $\bar{Q}_m$ is centred at $\bar{Q}_\infty$ and has variance B_∞/m, we have

$$(\bar{Q}_\infty \mid S_m, B_\infty) \sim (\bar{Q}_m, B_\infty/m). \tag{12}$$

As long as B_∞/m is small relative to $\bar{U}_\infty + B_\infty$, result (12) coupled with (10) gives

$$(Q \mid S_m, B_\infty) \sim t_{v_{\text{obs}}}(\bar{Q}_m, T_{\infty,m}), \tag{13}$$

where

$$T_{\infty,m} = \bar{U}_m + (1 + m^{-1})B_\infty = \bar{U}_\infty + (1 + m^{-1})B_\infty. \tag{14}$$

We are now prepared to integrate (13) over the posterior distribution of B_∞ given S_m. Since this integration cannot be done in closed form, we utilise some simple moment-matching approximations. For simplicity, the derivation is done for the scalar case but is appropriate when $\bar{U}_\infty$ is approximately proportional to T_∞. The key idea is to represent (13) as a mixture of normal posterior distributions over the distribution of the unknown $T_{\infty,m}$. That is, the distribution in (13) is equivalent to that obtained from

$$(Q \mid S_m, B_\infty, T_*) \sim N(\bar{Q}_m, T_*), \tag{15}$$

$$\left(\frac{T_{\infty,m}}{T_*} \,\middle|\, S_m, B_\infty\right) \sim \text{MS}_{v_{\text{obs}}}, \tag{16}$$

where $\text{MS}_{v_{\text{obs}}}$ is the mean square distribution on v_{obs} degrees of freedom, that is $\chi^2_{v_{\text{obs}}}/v_{\text{obs}}$, with

$$E\left(\frac{T_{\infty,m}}{T_*} \,\middle|\, S_m, B_\infty\right) = 1, \tag{17}$$

$$\text{var}\left(\frac{T_{\infty,m}}{T_*} \,\middle|\, S_m, B_\infty\right) = \frac{2}{v_{\text{obs}}}; \tag{18}$$

T_* is implicitly defined by (16), and (17) and (18) hold by the definition of a mean square distribution.

Now we need to integrate over $T_{\infty,m}$ and T_* in (15) and (16) given S_m, or equivalently integrate over B_∞ and T_* given S_m. We approximate this integral by finding the posterior mean and variance of the ratio of estimated to true variance, T_m/T_*, and matching these moments to a mean square distribution, whence we obtain an approximating t posterior distribution for Q.

From Rubin (1987, p. 91) we have the following approximation:

$$\left(\frac{T_m}{T_{\infty,m}} \,\middle|\, S_m\right) \sim \text{MS}_{v_m}, \tag{19}$$

where ν_m is given by (1) and T_m by (7). Hence, for the posterior mean of T_m/T_*, we have

$$E\left(\frac{T_m}{T_*}\,\middle|\,S_m\right)=E\left\{\frac{T_m}{T_{\infty,m}}\,E\left(\frac{T_{\infty,m}}{T_*}\,\middle|\,S_m,B_\infty\right)\,\middle|\,S_m\right\},$$

which from (17) and then (19) gives

$$E\left(\frac{T_m}{T_*}\,\middle|\,S_m\right)=E\left(\frac{T_m}{T_{\infty,m}}\,\middle|\,S_m\right)=1.$$

Next we calculate the posterior variance of T_m/T_*:

$$\operatorname{var}\left(\frac{T_m}{T_*}\,\middle|\,S_m\right)=E\left\{\operatorname{var}\left(\frac{T_m}{T_*}\,\middle|\,S_m,B_\infty\right)\,\middle|\,S_m\right\}+\operatorname{var}\left\{E\left(\frac{T_m}{T_*}\,\middle|\,S_m,B_\infty\right)\,\middle|\,S_m\right\},$$

which from (17) and (18) is

$$E\left\{\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{2}{\nu_{\text{obs}}}\,\middle|\,S_m\right\}+\operatorname{var}\left(\frac{T_m}{T_{\infty,m}}\,\middle|\,S_m\right),$$

which from (19) gives

$$\operatorname{var}\left(\frac{T_m}{T_*}\,\middle|\,S_m\right)=E\left\{\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{2}{\nu_{\text{obs}}}\,\middle|\,S_m\right\}+\frac{2}{\nu_m}. \tag{20}$$

Using the definition of ν_{obs} given in (11) and of $T_{\infty,m}$ given in (14), we obtain

$$E\left\{\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{2}{\nu_{\text{obs}}}\,\middle|\,S_m\right\}=2E\left\{\frac{T_m^2}{T_{\infty,m}}\frac{1}{\bar{U}_m\nu^*_{\text{com}}}-\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{B_\infty}{m\bar{U}_m\nu^*_{\text{com}}}\,\middle|\,S_m\right\},$$

which from (19) is

$$2\frac{T_m}{\bar{U}_m\nu^*_{\text{com}}}-2E\left\{\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{B_\infty}{m\bar{U}_m\nu^*_{\text{com}}}\,\middle|\,S_m\right\}. \tag{21}$$

A first-order Taylor-series expansion of the second term in (21) in terms of B_∞^{-1} gives

$$E\left\{\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{B_\infty}{m\bar{U}_m\nu^*_{\text{com}}}\,\middle|\,S_m\right\}\doteq\frac{B_m}{\bar{U}_m}\frac{1}{m\nu^*_{\text{com}}}. \tag{22}$$

Combining (21) and (22) yields, after some simple manipulations,

$$E\left\{\left(\frac{T_m}{T_{\infty,m}}\right)^2\frac{2}{\nu_{\text{obs}}}\,\middle|\,S_m\right\}\doteq\frac{2}{\nu^*_{\text{com}}(1-\hat{\gamma}_m)}\left(1-\frac{\hat{\gamma}_m}{m+1}\right), \tag{23}$$

where

$$\hat{\gamma}_m=\frac{(1+m^{-1})B_m}{T_m}.$$

Approximation (23) together with (20) gives

$$\operatorname{var}\left(\frac{T_m}{T_*}\,\middle|\,S_m\right)\doteq2\left(\frac{1}{\nu_m}+\frac{1}{\hat{\nu}_{\text{obs}}}C_m\right),$$

where $\hat{\nu}_{\text{obs}}$ is given in (5) and $C_m=1-\hat{\gamma}_m/(m+1)$.

Hence, assuming $(T_m/T_*\,|\,S_m)$ has a mean square distribution, its approximate degrees of freedom are

$$\left(\frac{1}{\nu_m}+\frac{1}{\hat{\nu}_{\text{obs}}}C_m\right)^{-1}.$$

Since C_m is typically close to one and setting $C_m = 1$ results in a conservative approximation, we take the degrees of freedom to be the harmonic total of v_m and $\hat{v}_{\text{obs}}$,

$$\tilde{v}_m = \left(\frac{1}{v_m} + \frac{1}{\hat{v}_{\text{obs}}}\right)^{-1},$$

thereby implying that

$$(Q \mid S_m) \sim t_{\tilde{v}_m}(\bar{Q}_m, T_m). \tag{24}$$

Under (24) and (9), the average fraction of information due to nonresponse is

$$\gamma_m = 1 - \text{tr}[\{\lambda(\tilde{v}_m)\bar{U}_m\}\{\lambda(v_{\text{com}})T_m\}^{-1}]/k, \tag{25}$$

where $\lambda(\gamma)$ is defined in (4). Result (25) is a generalisation of (3.3.17) in Rubin (1987).

5. Frequentist evaluation

Following the tradition of the methodological development of multiple imputation, in this section we report a small simulation study that assesses the frequentist validity of inferences based on (24), our modification of the standard repeated-imputation reference distribution.

The simulation study involved two variables, X and Y, which have a bivariate normal distribution, where X was always fully observed and Y was partially missing. The goal of the study was to assess and compare the frequentist performance of repeated-imputation confidence intervals for linear regression parameters constructed using the new repeated-imputation reference distribution (24) and the standard repeated-imputation reference distribution (8).

There were five factors in the simulation study; ρ, the correlation between X and Y, with levels 0·5 and 0·8; n, the complete-data sample size, with levels 10, 20 and 30; m, the number of imputations, with levels 3, 5 and 10; ϖ, the percent of missingness of Y, with levels 10, 20 and 30; and η, the slope parameter of the logistic missingness function given in (26), with levels -4, 0 and 4. The design of the study was a complete factorial design with 1000 replications for each of the 162 conditions.

Each replication of the simulation study consisted of four steps, as follows.

Step 1. Generate n independent and identically distributed draws, $(X_1, Y_1), \ldots, (X_n, Y_n)$, from a bivariate normal distribution with mean vector $\mu = (0, 0)^{\mathrm{T}}$ and variance–covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Step 2. Randomly choose ϖ percent of the observations on Y to be missing with probability

$$\propto \exp(\eta x_i^2)/\{1 + \exp(\eta x_i^2)\}. \tag{26}$$

Step 3. Draw m imputations of the missing observations from the posterior predictive distribution of the missing data assuming a bivariate normal distribution for the complete data, a flat prior distribution on the model parameters, that is $f(\mu, \Sigma) \propto |\Sigma|^{-3/2}$, and an ignorable missing data process.

Step 4. Obtain nominal 95% confidence intervals for $\beta_{Y|X} = \rho$ and $\beta_{X|Y} = \rho$ from the multiply imputed data separately under the standard repeated-imputation reference distribution (8) and the modified repeated-imputation reference distribution (24). For $\beta_{Y|X}$ and $\beta_{X|Y}$, the complete-data point estimate $\hat{Q}$ was the least-squares slope when regressing Y on X, $\hat{\beta}_{Y|X}$, and the least-squares slope when regressing X on Y, $\hat{\beta}_{X|Y}$, respectively. For each estimator, the complete-data variance estimate U was the usual estimate for a least-squares regression slope. The complete-data degrees of freedom, v_{com}, were $n - 2$.

Table 1 gives the main effect coverage results and indicates that, when the complete-data sample size is small, the average coverage of the modified repeated-imputation confidence-interval procedure is much closer to nominal than the standard repeated-imputation confidence-interval procedure, for all five simulation factors. In particular, even when the missing data mechanism is not missing completely at random, an assumption underlying the derivation of the modified procedure, i.e. when $\eta = -4$ or $\eta = 4$, the modified procedure has better average coverage than the standard. Not only did the modified procedure dominate the standard procedure with respect to the main effects of each of the factors, it did better for every simulation configuration.

Table 1. *Main effect averages of deviations from nominal coverage for* 95% *confidence intervals based on standard, equation* (8), *and modified, equation* (24), *reference distributions for estimands* $\beta_{Y|X}$ *and* $\beta_{X|Y}$. *Each entry is* 100 *times the average difference between estimated and nominal coverage over all simulation cells in which the factor given in the first column is equal to the corresponding level in the second column*

		Estimand $\beta_{Y\|X}$		Estimand $\beta_{X\|Y}$	
		Method		Method	
Factor	Level	Stand., eqn (8)	Modif., eqn (24)	Stand., eqn (8)	Modif., eqn (24)
ρ	0·5	−4·1	−0·9	−3·8	−0·9
	0·8	−3·8	−0·7	−3·4	−0·4
n	10	−6·3	−0·5	−6·2	−0·7
	20	−3·4	−1·1	−2·8	−0·7
	30	−2·2	−0·8	−1·9	−0·6
m	3	−3·8	−0·9	−3·5	−0·6
	5	−4·1	−0·9	−3·8	−0·8
	10	−4·0	−0·6	−3·5	−0·6
ϖ	10	−2·5	0·3	−2·3	0·3
	20	−3·9	−0·7	−3·5	−0·5
	30	−5·6	−2·0	−5·1	−1·8
η	−4	−3·0	−0·3	−3·3	−0·5
	0	−4·1	−0·8	−3·7	−0·7
	4	−4·8	−1·3	−3·8	−0·8

To assess how well the main effect results in Table 1 summarise the entire simulation, we calculated the residuals from predicting the coverage results under the main effects model. The dotplots, not shown here, of the residuals for each of the four combinations of estimand and interval procedure indicated that the main effect results given in Table 1 adequately summarise the simulation, especially for the modified procedure.

Based on the theoretical derivations and the simulation evidence presented here, we recommend that our modified repeated-imputation reference distribution be used in place of the standard repeated-imputation reference distribution in all analyses of multiply imputed data, especially with datasets having few primary sampling units.

Acknowledgement

The desirability of an adjustment to the multiple imputation degrees of freedom for small complete-data degrees of freedom was fostered by Statistical Solutions Ltd, and support for this work was partially provided by a grant from them and partially by the U.S. National Science Foundation.

References

Meng, X. L. (1994). Multiple imputation with uncongenial sources of input (with Discussion). *Statist. Sci.* **9**, 538–73.

Meng, X. L. & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–11.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–90.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years (with Discussion). *J. Am. Statist. Assoc.* **91**, 473–89.

Rubin, D. B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Assoc.* **81**, 366–74.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

Schafer, J. L., Khare, M. & Ezzati-Rice, T. M. (1993). Multiple imputation of missing data in NHANES III. In *Bureau of the Census Annual Research Conference*, Ed. M. Anderson-Brown, pp. 459–87. Washington: U.S. Dept. of Commerce.

Statistical Solutions (1997). *Solas for Missing Data Analysis 1.0.* Cork: Statistical Solutions Ltd.

[*Received February* 1998. *Revised January* 1999]