J. Statist. Comput. Simul., 2002, Vol. 72(4), pp. 309–318

Taylor & Francis
Taylor & Francis Group

A DEGREES-OF-FREEDOM APPROXIMATION IN MULTIPLE IMPUTATION

STUART R. LIPSITZ[a,*], MICHAEL PARZEN[b] and LUE PING ZHAO[c]

[a]*Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, 44 Binney Street, Boston MA 02115, U.S.A.;* [b]*Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637, U.S.A.;* [c]*Epidemiology Program, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, U.S.A.*

(Received 24 August 2001; Revised 28 September 2001)

When using multiple imputation to form confidence intervals with missing data, Rubin and Schenker (1986) proposed using a t-distribution with approximate degrees-of-freedom which is a function of the number of multiple imputations and the within and between imputation variance. In this t-approximation, Rubin and Schenker assume there are a finite number of multiple imputations, but an infinite number of observations in the sample. We propose a further degrees-of-freedom approximation which is a function of the within and between imputation variance, the number of multiple imputations, and the number of observations in the sample. When the number of observations in the sample is small, our approximate degrees-of-freedom may be more appropriate, as seen in our simulations.

Keywords: Bayesian bootstrap; Ignorable non-response

1 INTRODUCTION

In a given dataset, one usually has P variables on N independent subjects. Often, some subset of the P variables are missing for a subject. For example, the data in Table I, collected by the first author, gives housing prices (Y), age of the house ($X1$) and size of the house ($X2$) for 25 homes in the Dallas, TX area. There are $N = 25$ observations, but only 19 (76%) have both $Y, X1$ and $X2$ recorded; 6 observations have $X1$ missing. In this dataset, we are interested in the regression of Y on $X1$ and $X2$, so the problem becomes estimating a regression function when a covariate has some values missing.

Missing data poses many problems, including the practical fact that most computer packages only use the subjects with complete data ('complete cases'). Suppose, because of some underlying missing data mechanism, only n out of the N subjects have complete data. Following the nomenclature of Rubin (1976) and Little & Rubin (1987), a hierarchy of missing data mechanisms can be distinguished. First, if the missing data mechanism is independent of both $X = (\mathrm{X1}, \mathrm{X2})$ and Y, then the missing data is said to be *missing*

* Corresponding author.

ISSN 0094-9655 print; ISSN 1563-5163 online
DOI: 10.1080/00949650290007405

TABLE I Housing Price Data.

Price of house (Dollars)	*Age (Years)*	*Size (Sq feet)*
77900	–	1611
120100	9	2116
130300	5	2838
84400	4	1656
97500	–	2438
85300	–	1934
91500	8	1645
88500	5	1529
92800	4	1825
104300	–	2362
53700	8	1273
113000	5	2008
108300	–	2419
96800	6	2253
105400	10	1881
106400	3	2035
92900	–	2282
93800	6	1788
72200	7	1652
60400	4	1258
83100	4	1697
78200	5	1780
96500	6	1688
80700	9	1201
105100	9	1701

completely at random (MCAR). When the probability of non-response depends on the observed data, but not on the missing values, the missing data is said to be *missing at random* (MAR). Clearly, MCAR is a special case of MAR, and often no distinction is made between these two mechanisms and they are referred to as being *ignorable*. We caution, however, that the use of the term ignorable does not imply that the individuals with missing data can simply be ignored. Rather, the term ignorable is used here to indicate that it is not necessary to specify a model for the missing data mechanism in a likelihood-based analysis of the data. That is, the missing data mechanism can be ignored.

With ignorable non-response (Rubin, 1976), the EM-algorithm with a likelihood (Dempster *et al.*, 1977) can be used to obtain the maximum likelihood estimate. Another method for consistently estimating parameters with ignorable non-response is the method of multiple imputation (Rubin, 1978), which can be thought of as an approximate Monte Carlo EM-algorithm. The basic idea is to 'fill-in' the missing values with some 'appropriate' number to give a completed dataset, and then perform the usual analysis on this dataset. Using multiple imputation, we create two or more completed datasets, do the usual analysis on each completed dataset, then draw inferences based on both the within and between imputation variability. The key step in Rubin's (1978) multiple imputation is 'filling-in' the missing data by drawing from the conditional distribution of the missing data given the observed data. This usually entails posing a parametric or semi-parametric model for the joint distribution of X and Y and using it to derive the conditional distribution of the missing data (either X or Y, whichever is missing) given the observed data (either X or Y whichever is observed). This is exactly what is done in the E-step of the EM-algorithm, and, thus Multiple Imputation can be considered a Monte Carlo EM-algorithm. As such, the variance of the multiple imputation estimate should be approximately the same (depending on the imputation method) as the maximum likelihood estimate from the EM-algorithm.

When using multiple imputation to form confidence intervals for parameters when data are missing, Rubin and Schenker (1986) proposed using a *t*-distribution with degrees-of-freedom which is a function of the within and between imputation variance and the number of multiple imputations. In this *t*-approximation, Rubin and Schenker assume that there are a finite number of multiple imputations, but an infinite number of observations in the sample. For the data in Table I, with only 18 subjects, and 6 missing, the number of observations is not large, so that Rubin and Schenker's degrees-of-freedom approximation may not be accurate. We propose a degrees-of-freedom approximation which is a function of the within and between imputation variance, the number of multiple imputations, and the number of observations in the sample. The degrees-of-freedom approximation can be used with any statistical method (ordinary least squares regression, logistic regression, generalized linear models) in which multiple imputation is used. Section 2 discusses multiple imputation in general and Section 3 describes the degrees-of-freedom approximation. In Section 4, we apply the degrees-of-freedom approximation to the apple data in Table I. Section 5 gives the results of simulations comparing our degrees-of-freedom approximation to Rubin and Schenker's approximation in the context of missing data in a one-sample problem.

2 MULTIPLE IMPUTATION

Suppose we have a sample of N independent $P \times 1$ random vectors $\{\mathbf{Y}_i\}$. With missing data, we can partition $\mathbf{Y}_i$ into two parts, $\mathbf{Y}_{i,\text{obs}}$ and $\mathbf{Y}_{i,\text{mis}}$, where $\mathbf{Y}_{i,\text{obs}}$ contains the observed variables and $\mathbf{Y}_{i,\text{mis}}$ contains the missing variables on subject i. Further, we let

$$\mathbf{Y}'_{\text{obs}} = (\mathbf{Y}'_{1,\text{obs}}, \ldots, \mathbf{Y}'_{N,\text{obs}}) \text{ and } \mathbf{Y}'_{\text{mis}} = (\mathbf{Y}'_{1,\text{mis}}, \ldots, \mathbf{Y}'_{N,\text{mis}}).$$

Our goal is estimate $\boldsymbol{\beta}$, the parameter vector of interest. We will use multiple imputation to 'fill-in' $\mathbf{Y}_{\text{mis}}$ and then estimate $\boldsymbol{\beta}$.

For a detailed summary of multiple imputation see Rubin and Schenker (1986) and Rubin (1987). We briefly review some relevant parts. Given that we have imputed $\mathbf{Y}_{\text{mis}}$ as $\mathbf{Y}^*_{\text{mis}}$ by one of the methods given in Rubin and Schenker (1986), we calculate the estimate of the parameter of interest, and its estimated variance, using the completed data $(\mathbf{Y}_{\text{obs}}, \mathbf{Y}^*_{\text{mis}})$. In particular, for the method of choice (*i.e.*, maximum likelihood, method of moments), we calculate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y}) = \hat{\boldsymbol{\beta}}(\mathbf{Y}_{\text{obs}}, \mathbf{Y}^*_{\text{mis}})$ and the within imputation variance $U = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$. We then independently impute $\mathbf{Y}_{\text{mis}}$ a large number of M times. The M completed datasets give us $\hat{\boldsymbol{\beta}}^m$ and U^m, for $m = 1, \ldots, M$.

With no missing data, suppose that inference about the parameter vector $\boldsymbol{\beta}$ is made by

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim N(0, U).$$

With M imputations, the multiple imputation estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^* = \frac{\sum_{m=1}^{M} \hat{\boldsymbol{\beta}}^m}{M}.$$

Further, when N and M are large, one can make normal based inferences for $\boldsymbol{\beta}$ with (Rubin, 1978),

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^*) \sim N(0, V),$$

where

$$V = \widehat{W} + \left(\frac{M+1}{M}\right)\widehat{B}, \tag{1}$$

$$\widehat{W} = \frac{\sum_{m=1}^{M} U^m}{M}$$

is the average within imputation variance, and

$$\widehat{B} = \frac{\sum_{m=1}^{M}(\hat{\boldsymbol{\beta}}^m - \hat{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}^m - \hat{\boldsymbol{\beta}}^*)'}{M-1}$$

is the between imputation variance.

3 THE DEGREES-OF-FREEDOM APPROXIMATIONS

Suppose β is a scalar. When one uses a small number of multiple imputations (M), Rubin and Schenker (1986) proposed assuming

$$V^{-1/2}(\beta - \hat{\boldsymbol{\beta}}^*) \sim t_{df_R},$$

a t-distribution with degrees-of-freedom equal to

$$df_R = \frac{[\widehat{W} + ((M+1)/M)\widehat{B}]^2}{((M+1)/M)^2\widehat{B}^2/(M-1)} = \left[1 + \left(\frac{M}{M+1}\right)\frac{\widehat{W}}{\widehat{B}}\right]^2 (M-1), \tag{2}$$

where $\widehat{W}$ and $\widehat{B}$ are defined in Section 2.

Instead, we propose assuming

$$V^{-1/2}(\beta - \hat{\boldsymbol{\beta}}^*) \sim t_{df},$$

where

$$df = \frac{[\widehat{W} + ((M+1)/M)\widehat{B}]^2}{\widehat{W}^2/(N-1) + ((M+1)/M)^2\widehat{B}^2/(M-1)}. \tag{3}$$

The approximate degrees-of-freedom in (3) was obtained by applying the Satterthwaite approximation (1946) to V in (1), and assuming that $\widehat{W}$ and $\widehat{B}$ are independent mean squares, with degrees-of-freedom -1) and (M $-$ 1), respectively. The degrees-of-freedom approximation we propose in (3) does not equal (2) proposed by Rubin and Schenker (1986), unless $N \to \infty$. Our degrees-of-freedom approximation in (3) takes the sample size (N) into account, whereas Rubin and Schenker's in (2) does not.

Note that our approximation (3) is always less than or equal to Rubin and Schenker's (2), and thus will always give more conservative (wider) confidence intervals. Now, suppose there is no missing data so that $\widehat{B} = 0$ and $\hat{\beta}$ is the sample mean. Then Rubin and Schenker's approximation (2) equals ∞, implying we would be using a normal quantile (t with ∞ degrees-of-freedom) for confidence intervals for β. However, our approximation (3) equals

$(N - 1)$, which is the degrees-of-freedom ordinarily used with a t-statistic for a complete data sample mean when forming confidence intervals. Thus, with no missing data, unlike Rubin and Schenker's approximation, our approximation reduces to what is commonly used as a degrees-of-freedom approximation.

When $\boldsymbol{\beta}$ is a $K \times 1$ vector, and one wants a confidence interval for the kth element of $\boldsymbol{\beta}$, we suggest using the t-distribution with approximate degrees-of-freedom

$$df_k = \frac{[\widehat{W}_k + ((M+1)/M)\widehat{B}_k]^2}{\widehat{W}_k^2/(N-K) + ((M+1)/M)^2\widehat{B}_k^2/(M-1)}, \tag{4}$$

where $\widehat{W}_k$ and $\widehat{B}_k$ are the kth diagonal elements of $\widehat{W}$ and $\widehat{B}$, respectively. The degrees-of-freedom approximation in (4) was obtained using the Satterthwaite approximation (1946), assuming that $\widehat{W}_k$ and $\widehat{B}_k$ are independent mean squares, with degrees-of-freedom $(N - K)$ and $(M - 1)$, respectively. Suppose there is no missing data ($\widehat{B} = 0$) and we have a linear regression model with parameters estimated by ordinary least squares. Then (4) reduces to $(N - K)$, the usual degrees-of-freedom for ordinary least squares. Thus, again, with no missing data, our approximation reduces to what is commonly used as a degrees-of-freedom approximation.

Using the degrees-of-freedom approximation in (4), one gets a different degrees-of-freedom for each element of $\boldsymbol{\beta}$. Alternatively, one can extend the ideas in Rubin (1987, page 78), to get a single degrees-of-freedom approximation. First, note that we can rewrite (4) as

$$df_k = \frac{[(\widehat{B}_k/\widehat{W}_k)^{-1} + ((M+1)/M)]^2}{1/(N-K)(\widehat{B}_k/\widehat{W}_k)^{-2} + 1/(M-1)((M+1)/M)^2} \tag{5}$$

To get a single degrees-of-freedom approximations, in (5), we replace $\widehat{B}_k/\widehat{W}_k$ with $tr(\widehat{\mathbf{B}}\widehat{\mathbf{W}}^{-1})/K$, where $tr(\cdot)$ is the trace of a matrix. However, because (4) better reflects the information in the data for making inferences about β_k, in the rest of this paper, we use the approximation given in (4).

4 EXAMPLE: HOUSING PRICE DATA

In Table I we have data on 25 houses sold in the Dallas metropolitan area in 1990. This data was collected by the first author from a random sample of the Sunday edition of the *Dallas Morning News*. The response variable of interest is the selling price of the house (price), measured in hundreds of thousands of dollars. The covariates are the size of the house (*Size*), measured in units of a thousand square feet of heated floor space, and the age of the house in years (age). We are primarily interested in building a model to explain the mean log sale price in terms of these two covariates. All variables are fully observed except for the variable age, which is missing for 6 of the 25 observations. The regression equation we will be fitting is

$$\log(SalesPrice) = \beta_{age}(Age) + \beta_{size} \times (Size) + \varepsilon,$$

where ε is the usual standard normal error term.

The missing values of the variable *Size* are imputed using the Splus NORM routines provided by Schafer (1997). These routines perform maximum-likelihood estimation on

TABLE II Estimates for the Housing Data.

Parameter	*Estimate*	*Estimated standard error*	*Degrees-of-freedom*		*95% Confidence interval*	
			new[a]	*RS*[b]	*new*[a]	*RS*[b]
M = 5 *imputations*						
β_{age}	0.55	0.14	27.95	191.29	[0.26, 0.84]	[0.27, 0.83]
β_{size}	4.22	0.49	26.62	80.34	[3.22, 5.22]	[3.25, 5.19]
M = 20 *imputations*						
β_{age}	0.53	0.15	31.34	984.38	[0.23, 0.83]	[0.24, 0.82]
β_{size}	4.27	0.51	34.32	514.71	[3.24, 5.30]	[3.28, 5.27]
Ordinary least squares with complete cases						
β_{age}	0.53	0.16	17[c]	–	[0.19, 0.86]	–
β_{size}	4.43	0.56	17	–	[3.25, 5.62]	–

[a] 'new' is our approximation (3).
[b] RS is Rubin and Schenker's approximation (2).
[c] usual degrees-of-freedom with regression.

the matrix of incomplete data using the EM-algorithm. The data and Splus code for this example may be found at `http://gsb.uchicago.edu/fac/michael.parzen`.

After filling in the missing values of *Size* in the *m*th imputation, we estimate the parameters $(\beta_{\text{age}}, \beta_{\text{size}})$ using ordinary least squares.

The results are shown in Table II for $M = 5$ and $M = 20$ imputations. Note that our estimates of degrees of freedom (approximately $df_{\beta_{age}} = 28$ and $df_{\beta_{size}} = 27$ when $M = 5$ and approximately $df_{\beta_{age}} = 31$ and $df_{\beta_{size}} = 34$ when $M = 20$) are much smaller than Rubin and Schenker's (both greater than 80 when $M = 5$ and both greater than 500 when $M = 20$). In particular, for all intents, Rubin and Schenker's degrees of freedom approximation suggests we use essentially a normal approximation in both imputations. Our degrees of freedom approximation leads to more conservative 95% confidence intervals, which would appear to be more appropriate given the small sample size (N = 25). Table II also gives estimates and confidence intervals when applying ordinary least squares to the complete cases (the 19 observations in which all the data are observed).

In Table II, the estimates of (β_{age}) are similar for the three methods (between 0.53 and 0.55), but the estimate of (β_{size}) increases slightly in size from 4.22 when $M = 5$ to 4.27 when $M = 20$ and then increases to 4.43 for the complete case method. This could be due to greater variablity in the *Size* variable. From Table II, one also sees that the estimated standard errors using the complete cases are the largest, and the estimated standard errors using multiple imputation are smaller than those from complete cases. In general, multiple imputation can be considered an approximate Monte Carlo EM-algorithm. In particular, if one uses a fully normal method of multiple imputation (Rubin, 1987), and (X_i, Y_i) are truly multivariate normal, then the variance of the MLE and the multiple imputation estimate will be the same (asymptotically), as well as asymptotically efficient. Thus, it is not surprising that multiple imputation has smaller estimated variance than complete cases.

5 SIMULATION AND DISCUSSION

To compare the performance of the degrees-of-freedom approximation given in (2) and Rubin and Schenker's given in (3), we used simulation to look at the coverage probability of confidence intervals for the mean for a simple one sample problem. We

generated i.i.d. random variables $\{Y_i\}$ (i = 1, . . . , N), and we want a confidence interval for $\mu = E(\mathrm{Y_i})$. Each Y_i is independently missing with probability f (this is an ignorable non-response model and thus multiple imputation is appropriate). The simulation was done in Splus and the code may be obtained at `http : \\gsb.uchicago.edu\fac\michael.parzen\research`. We note that random number generation in Splus is adapted from Marsaglia (1973).

We performed 96 sets of simulations corresponding to the cross-classification of the factors:

A. *Distribution:* $N(0, 1)$; lognormal $[= \exp(\mathrm{N}(0, 1))]$ with mean $e^{1/2} \approx 1.65$ and variance $e(\mathrm{e} - 1) \approx 4.67$; and standard Laplace with mean 0 and variance 2.
B. *Nominal level of interval:* 90%, 95%.
C. *Sample size* (N): 10, 15, 20, 30
D. *Fraction missing* (f): 0.1, 0.2, 0.4, 0.6

We fixed the number of multiple imputations at $M = 2$. For computational simplicity, we only considered the Approximate Bayesian Bootstrap method of multiple imputation as described in Rubin and Schenker (1986). In particular, suppose n out of the N subjects are observed; using the approximate bayesian bootstrap, for the mth imputation:

1. Independently draw n observations from the observed y_i's with probabilities $1/n$.
2. Draw the $N - n$ missing observations with replacement from the n values drawn in step 1.
3. Calculate $\bar{Y}^m$.

We chose this imputation method because it is a 'non-parametric' method of multiple imputation, in that it is appropriate regardless of the distribution of the underlying Y_i's. As discussed later, other imputation methods for these data may be more appropriate if one is willing to make further assumptions about the distribution of Y_i (such as imputation based on the normal distribution when assuming the data are normal). However, we are more interested in the relative performance of the degrees-of-freedom approximations, so that the method of imputation is not the critical issue in this paper. Further, in reality, with univariate Y_i and ignorable non-response, the most efficient method is to discard (N − n) missing observations, and obtain the MLE using only the n observed Y_i's. For example, for most distributions, the sample mean of the n non-missing Y_i's is the MLE of the population mean. It might appear that, by imputing the (N − n) non-missing Y_i's, we are arbitrarily adding (N − n) observations, and can thus make the variance of the multiple imputation estimate arbitrarily small. However, recall from Eq. (1), that the variance of the resulting multiple imputation estimate is

$$V = \widehat{W} + \left(\frac{M+1}{M}\right)\widehat{B}.$$

It is true that $\widehat{W}$, the average of the within imputation variance (basically, the variance of the estimate from a sample of size N) will be smaller than variance of the estimate using only the n non-missing Y_i's. However, the variance of the multiple imputation also has the extra piece $((\mathrm{M} + 1)/\mathrm{M})\widehat{\mathrm{B}}$, corresponding to the between imputation variance. One can show that the multiple imputation variance V can be no smaller than the variance of the estimate using only the n non-missing Y_i's. Even though we would really only need to use the n non-missing Y_i's, we again are using this simple univariate setting to compare the degrees-of-freedom approximations.

TABLE III Simulated Coverages of Approximate Bayesian Bootstrap with $M = 2$ imputations, with nominal levels 90% and 95%, fraction missing f, and three different underlying distributions (Normal, Laplace, Lognormal).

			Normal		*Laplace*		*Lognormal*	
N	*f*	*df**	*90%*	*95%*	*90%*	*95%*	*90%*	*95%*
10	0.1	RS	75.8	80.4	76.5	82.2	68.5	73.1
		new	88.6	93.8	89.0	95.6	78.1	83.3
	0.2	RS	81.5	87.9	82.4	88.7	73.6	78.5
		new	86.9	93.9	87.3	94.3	77.5	82.9
	0.4	RS	79.9	86.2	83.3	89.0	70.3	74.8
		new	83.4	89.4	86.1	91.7	72.8	77.7
	0.6	RS	75.1	81.9	76.1	82.4	65.7	71.4
		new	78.8	84.5	78.4	85.2	67.1	73.5
15	0.07	RS	81.5	86.9	81.8	86.6	74.8	76.3
		new	89.3	94.1	90.4	95.0	79.0	84.4
	0.2	RS	87.1	92.1	85.8	91.1	77.2	82.3
		new	89.1	94.4	88.2	93.6	79.1	83.9
	0.4	RS	84.1	89.6	83.5	89.7	74.5	80.7
		new	85.4	91.2	84.7	91.1	76.6	82.5
	0.6	RS	78.9	84.3	80.8	85.7	69.6	76.9
		new	79.8	85.5	81.4	86.9	70.9	77.9
20	0.1	RS	86.6	91.9	88.1	93.2	78.1	83.8
		new	88.7	94.8	89.7	95.2	79.8	85.9
	0.2	RS	86.7	92.2	85.8	91.9	81.6	85.9
		new	87.8	93.3	87.0	93.4	82.8	86.8
	0.4	RS	85.9	90.5	84.9	90.5	73.7	79.6
		new	86.8	92.0	85.9	91.8	74.7	80.3
	0.6	RS	80.6	87.6	82.1	87.1	73.7	79.6
		new	81.16	88.8	82.9	88.2	74.3	80.6
30	0.1	RS	88.3	93.5	89.5	94.6	82.2	87.1
		new	89.1	94.4	90.4	95.2	83.1	87.9
	0.2	RS	88.3	93.3	87.9	94.1	80.4	86.1
		new	89.5	94.0	88.9	94.8	81.2	87.1
	0.4	RS	85.9	91.1	85.4	91.3	78.5	84.8
		new	86.6	91.8	85.9	92.3	79.1	85.5
	0.6	RS	82.1	87.8	83.5	84.4	75.4	81.2
		new	82.3	88.5	83.8	89.4	75.7	81.9

*RS is Rubin and Schenker's degrees-of-freedom approximation in (2); new is our proposed degrees-of-freedom approximation in (3).

In each simulation set, we used 2000 replications. For a given replication, the two-sided $(1 - \alpha)100\%$ confidence interval is calculated by

$$\bar{Y}^* \pm t_{\alpha/2,v}\sqrt{\hat{V}},$$

where $\bar{Y}^*$ is the average of the sample means over the multiple imputations and $t_{\alpha/2,v}$ is the upper $(1 - \alpha/2)$ quantile of the central t-distribution with v degrees-of-freedom (which equals either (2) or (3]). The coverage probability in Table I is calculated as the proportion of replications in which the confidence interval contains the true mean.

First, in Table I, before comparing the degrees-of-freedom approximations, we see the same general trends found in Rubin and Schenker (1986). In particular, as N increases, the coverage is closer to the nominal level; as the fraction (f) of missing data decreases,

the coverage is closer to the nominal level. Further, the coverage is closer to the nominal level for the symmetric distributions (Normal, Laplace) than the skewed lognormal distribution.

Next, we compare the degrees-of-freedom approximations as N changes, the distribution changes, and the fraction of missing data (f) changes. From Table I, we see that our degrees-of-freedom approximation (3) always leads to confidence intervals closer to the nominal level than Rubin and Schenker's in (2]. As expected, as N gets larger, the difference in the coverage between (3) and (2) becomes smaller since (3) and (2) become almost identical. However, when N is small (≤ 15), our approximation in (3) is more appropriate. For example, when $N = 10$, there are 2 missing observations (f $= 0.2$), and the true distribution is Laplace, the coverage for 95% confidence intervals increases from 88.7% for Rubin and Schenker's approximation to 94.3% using our approximation.

As the fraction (f) of missing data increases, our approximation and Rubin and Schenker's approximation become more similar; this is apparently because our degrees-of-freedom in (3) is dominated by the between imputation variance when there is a lot of missing data (and thus the term $\widehat{W}^2/(\mathrm{N}-1)$ in (3) has very little impact). For example, when $N = 20$, the distribution is normal, and $f = 0.1$, the coverage probability for a 95% interval increases from 91.9% when using Rubin and Schenker's approximation to 94.8% when using ours, a 2.9% difference. However, the fraction of missing data increases to $f = 0.6$ (with $N = 20$ and a true normal distribution), a 95% confidence interval using (2) gives coverage 87.6%, and (3) gives coverage 88.8%, a 1.2% difference.

The poor coverage that is seen as the fraction of missingness (f) gets larger is due to the method of imputation (Approximate Bayesian Bootstrap), which is a 'non-parametric' type of imputation. In simulations not shown, when we imputed the data using a normal distribution method of imputation (Rubin and Schenker, 1986), the coverage in the first column of Table I (in which the data are truly normal) looks much better. In fact, if we used a normal method of imputation, coverage for 95% confidence intervals for most entries in the first column of Table I is about 93.5% using Rubin and Schenker's approximation, and about 94.3% using our approximation. Unfortunately, though, if we impute using a normal distribution method, the coverage in the last two columns look much worse than with the Approximate Bayesian Bootstrap. Similarly, if we impute with a log-normal distribution method, then the last column looks much better, but the first two columns look much worse than with the Approximate Bayesian Bootstrap. We chose the Approximate Bayesian Bootstrap because it should be robust for any distribution, although we see the coverage is still low for many entries in the table.

Because of the broad range of possible data, missing data models, imputation methods and parameters of interest (*e.g.*, median, mean, regression parameters), it is difficult to draw definitive conclusions from a simulation. One can only make general suggestions. Again, we have looked at a simple case to study the properties of the degrees of freedom approximation. We encourage using our degrees-of-freedom approximation (3], since, when N is large, it is almost identical to Rubin and Schenker's (2], and, when N is small, it appears more appropriate in our limited simulations, and can increase the coverage probability by up to 14%.

Acknowledgements

The authors are grateful for the support provided by the following grants from the United States' Institutes of Health: HL 69800, AHRQ 10871, HL52329, HL61769, CA 70101 and the IBM Corporation Faculty Research Fund at the Graduate School of Business, University of Chicago.

References

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

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

Marsaglia, G. (1973). *Random Number Package: "Super-Duper"*. School of Computer Science, McGill University.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys – a phenominological Bayesian approach to nonresponse. In: *Proceedings of the International Statistical Institute*, Manila, pp. 517–532.

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

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366–374.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110–114.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*. The Iowa State University Press, Ames, Iowa.