

Statistical Practice

A Potential for Bias When Rounding in Multiple Imputation

Nicholas J. HORTON, Stuart R. LIPSITZ, and Michael PARZEN

With the advent of general purpose packages that support multiple imputation for analyzing datasets with missing data (e.g., Solas, SAS PROC MI, and S-Plus 6.0), we expect much greater use of multiple imputation in the future. For simplicity, some imputation packages assume the joint distribution of the variables in the multiple imputation model is multivariate normal, and impute the missing data from the conditional normal distribution for the missing data given the observed data. If the possibly missing data are not multivariate normal (say, binary), imputing a normal random variable can yield implausible values. To circumvent this problem, a number of methods have been developed, including rounding the imputed normal to the closest observed value in the dataset. We show that this rounding can cause biased estimates of parameters, whereas if the imputed value is not rounded, no bias would occur. This article shows that rounding should not be used indiscriminately, and thus some caution should be exercised when rounding imputed values, particularly for dichotomous variables.

KEY WORDS: Fully normal imputation; Missing completely at random; Missing data.

1. INTRODUCTION

Missing data is a common occurrence in most datasets. An increasingly popular method for parameter estimation with missing data is the method of multiple imputation (Rubin 1978). The basic idea is to “fill-in” the missing values with some “appropriate” value to give a completed dataset, and perform a complete data analysis. Using multiple imputation, the user creates two or more completed datasets, carries out the analysis on each completed dataset, and draws inferences based on both the within and between imputation variability. We expect the advent of general purpose packages that support multiple imputation, such as Solas (2001), SAS PROC MI (2001), and S-Plus 6.0 (Schimert

et al. 2001) will lead to much greater use of multiple imputation in the future.

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 positing a parametric model for the data and using it to derive the conditional distribution of the missing data given the observed data. A detailed summary of multiple imputation can be found in Rubin and Schenker (1986) and Rubin (1987). Some imputation packages assume the joint distribution of the variables in the dataset is multivariate normal. In this case, assuming the data are missing at random (Little and Rubin 1987), the conditional distribution of the missing data given the observed data will also be multivariate normal. This approach is attractive because the conditional multivariate normal distribution is easy to sample from.

However, if the possibly missing data are not normal (say, binary), imputing a normal random variable will give an implausible value. An approach to circumvent this problem suggests sampling a value from the conditional distribution of the missing data given the observed data, then creating an imputed value by finding the closest observed value in the dataset (Rubin 1987, p. 168, example 5.2). For discrete variables, this is equivalent to rounding to the closest observed value in the dataset. This approach is most appropriate when the missing data take on many values, and the marginal distribution is approximately unimodal and symmetric (Schafer 1997, p. 148). However, the approach has been proposed in settings where these assumptions are not tenable, such as with dichotomous variables (e.g., Schafer 1997, p. 148; Sinharay, Stern, and Russell 2001). This may tempt analysts to round discrete random variables with distributions that are clearly nonnormal. For example, if the missing value can only take values 0 and 1, but it is imputed based on a standard normal distribution, the analyst would round the imputed value to 1 if the sampled normal variate is greater than or equal to 0.5, and round the imputed value to 0 if the sampled normal variate is less than 0.5.

A better approach would be to posit a model for a discrete missing value. Rubin (1987 p. 169, example 5.3) described a logistic regression imputation model for a dichotomous outcome, and detailed the steps for imputation. Unfortunately, this type of model is not supported by all multiple imputation implementations, and analysts may be tempted to use less appropriate methods.

This article cautions against the use of rounding to impute a missing value of a discrete variable, because it can introduce bias in the parameter estimate of interest. The following section shows that, for a single sample of N Bernoulli observations, imputing based on the normal distribution and rounding can give

Nicholas J. Horton is Assistant Professor, Department of Mathematics, Smith College, Northampton, MA 01063 (E-mail: nhorton@smith.edu). Stuart R. Lipsitz is Professor, Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC 29425. Michael Parzen is Associate Professor, Department of Decision and Information Analysis, Goizueta Business School, Emory University, Atlanta, GA 30322. The authors are grateful to Ken Kleinman for comments on an earlier draft of the article, to the reviewers and associate editor for useful suggestions, and for the support provided by NIH grants MH54693, HL69800, HL52329, HL61769, GM29745, CA70101, and AHRQ10871.

a biased estimate of the Bernoulli probability of success. However, if the imputed normal variate is not rounded, the estimated Bernoulli probability will be unbiased.

2. AN EXAMPLE OF BIAS FROM ROUNDING

Suppose we intend to sample N , iid Bernoulli random variables, $\{Y_1, \dots, Y_N\}$, where $p = E(Y_i)$ is the probability of success. Unfortunately, suppose we are only able to obtain n out of the N Bernoulli observations. Without loss of generality, we assume the first “ n ” data points constitute the observed data, $\{Y_1, \dots, Y_n\}$, and the last $(N - n)$ data points constitute the missing data, $\{Y_{n+1}, \dots, Y_N\}$. Further, we assume the missing data are missing completely at random, so that $\{Y_1, \dots, Y_n\}$ is a completely random sample of n iid Bernoulli random variables, each with probability p of success. Our interest lies in estimating p . In this case, the minimum variance unbiased estimate (MVUE) of p is just the sample mean of the observed data, which we denote by \hat{p} , that is,

$$\hat{p} = n^{-1} \sum_{i=1}^n Y_i.$$

Even though we have the MVUE, Rubin and Schenker (1986) have used this simple missing data example to illustrate different methods of imputation, and we use it here to illustrate the bias that can be encountered by “rounding” when imputing. Rubin (1978) proposed using imputation to “fill-in” the missing data $\{Y_{n+1}, \dots, Y_N\}$ using the observed data $\{Y_1, \dots, Y_n\}$, and then using the “filled-in” data to estimate p , as described in the following.

Suppose we use a fully normal (FN) imputation method (Rubin and Schenker 1986) to impute $\{Y_{n+1}, \dots, Y_N\}$. In the FN method, we assume that the Y_i are iid from a normal distribution with mean p and variance σ^2 , that is, $\mathcal{N}(p, \sigma^2)$. We denote the sample variance of the observed Y_i 's by

$$s^2 = (n - 1)^{-1} \sum_{i=1}^n (Y_i - \hat{p})^2 = (n - 1)^{-1} n \hat{p}(1 - \hat{p}).$$

The FN method first draws a value of σ^2 , say σ^{*2} , from a $(n - 1)s^2/\chi_{n-1}^2$ distribution, where χ_{n-1}^2 is a chi-square random variable with $(n - 1)$ degrees of freedom; then a value of p , say p^* , is drawn from $\mathcal{N}(\hat{p}, \sigma^{*2}/n)$. Then, the filled-in values for the missing data, say $(Y_{n+1}^*, \dots, Y_N^*)$, are drawn as iid $\mathcal{N}(p^*, \sigma^{*2})$. Using this filled-in data, the imputation estimate of p is

$$\hat{p}^* = N^{-1} \left[\sum_{i=1}^n Y_i + \sum_{i=n+1}^N Y_i^* \right]. \quad (1)$$

Using multiple imputations we construct M filled-in datasets (where M is often small, e.g., 5 or 10). For our purposes, we will look at the bias in the estimate (1) using the FN method, as well as the bias in the estimate using FN imputation with rounding (RND). By RND we mean that, the imputed value for Y_i ($i > n$),

$$\mathcal{Y}_i^* = \begin{cases} 1 & \text{if } Y_i^* \geq 0.5 \\ 0 & \text{if } Y_i^* < 0.5 \end{cases}, \quad (2)$$

where Y_i^* is obtained using the FN method. Now, we look at the biases $E[Y_i^* - p]$ and $E[\mathcal{Y}_i^* - p]$.

First, consider FN imputation (without rounding). One can show the $E[Y_i^*] = p$ using conditional expectations. Under the FN imputation model, Y_i^* is drawn from a $\mathcal{N}(p^*, \sigma^{*2})$ after p^* is drawn from $\mathcal{N}(\hat{p}, \sigma^{*2}/n)$. The expectation of Y_i^* under the FN imputation model, conditional on σ^{*2} , is given by

$$\begin{aligned} E_{\sigma^{*2}}[Y_i^*] &= E_{\hat{p}|\sigma^{*2}}\{E_{p^*|\sigma^{*2}}[Y_i^*]\} \\ &= E_{\hat{p}|\sigma^{*2}}\{p^*\} \\ &= \hat{p}. \end{aligned}$$

Then, because $E_{\sigma^{*2}}[Y_i^*] = \hat{p}$ and $E[\hat{p}] = p$, it follows that $E[Y_i^*] = p$. Thus, the FN imputation without rounding method, which incorrectly assumes a normal distribution that can yield “implausible values,” nevertheless will produce an unbiased estimate of p .

Now, consider FN imputation with rounding (RND), with imputed value \mathcal{Y}_i^* and with expected value

$$E[\mathcal{Y}_i^*] = \Pr[\mathcal{Y}_i^* = 1] = \Pr[Y_i^* > 0.5].$$

To obtain this probability, we first rewrite Y_i^* in terms of \hat{p} , which entails some probability arguments, as follows. In terms of (μ^*, σ^{*2}) , $Y_i^* = \mu^* + \sigma^* Z_1$, where $Z_1 \sim \mathcal{N}(0, 1)$. Next, because $\mu^* = \hat{p} + \frac{\sigma^*}{\sqrt{n}} Z_2$, where $Z_2 \sim \mathcal{N}(0, 1)$, and Z_2 is independent of Z_1 , we write

$$Y_i^* = \hat{p} + \frac{\sigma^*}{\sqrt{n}} Z_2 + \sigma^* Z_1 = \hat{p} + \sigma^* \left(Z_1 + \frac{Z_2}{\sqrt{n}} \right).$$

However,

$$Z = \sqrt{\frac{n}{n+1}} \left(Z_1 + \frac{Z_2}{\sqrt{n}} \right) \sim \mathcal{N}(0, 1),$$

so that

$$Y_i^* = \hat{p} + Z \sigma^* \sqrt{\frac{n+1}{n}}.$$

Next, from the FN imputation scheme, for a random variable $X \sim \chi_{n-1}^2$:

$$\begin{aligned} \sigma^{*2} &= \frac{(n-1)s^2}{X} = \frac{(n-1)[(n-1)^{-1}n\hat{p}(1-\hat{p})]}{X} \\ &= \frac{(n-1)^{-1}n\hat{p}(1-\hat{p})}{X/(n-1)}, \end{aligned}$$

so that

$$\begin{aligned} Y_i^* &= \hat{p} + \sqrt{\frac{(n+1)\hat{p}(1-\hat{p})}{(n-1)}} \left(\frac{Z}{\sqrt{X/(n-1)}} \right) \\ &= \hat{p} + \sqrt{\frac{(n+1)\hat{p}(1-\hat{p})}{(n-1)}} T, \end{aligned}$$

where $T \sim t_{(n-1)}$ (a t -random variable with $(n - 1)$ degrees of freedom). Then, under the FN imputation model, conditional on \hat{p} ,

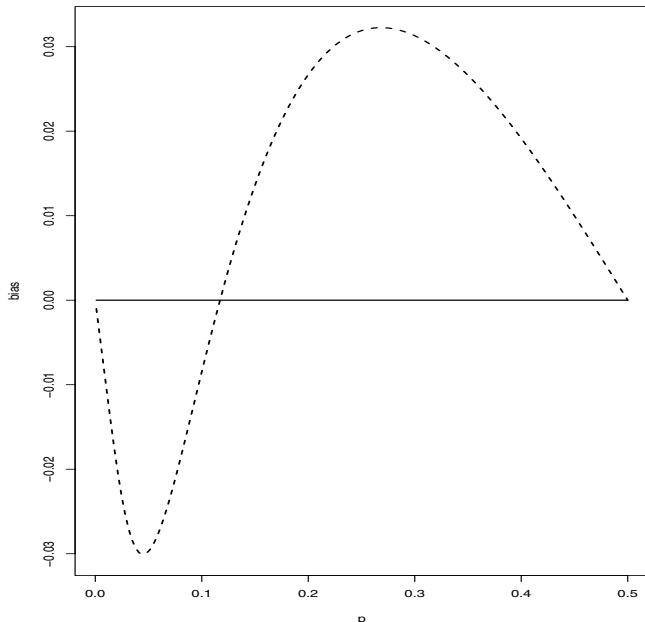


Figure 1. Bias ($E[\mathcal{Y}_i^*] - p$) of RND imputation for various probabilities.

$$\begin{aligned} \Pr[Y_i^* > 0.5 \mid \hat{p}] &= \Pr[\hat{p} + \sqrt{(n+1)\hat{p}(1-\hat{p})/(n-1)}T > 0.5 \mid \hat{p}] \\ &= \Pr\left[T > \frac{0.5 - \hat{p}}{\sqrt{(n+1)\hat{p}(1-\hat{p})/(n-1)}} \mid \hat{p}\right] \\ &= 1 - F_{t(n-1)}\left(\frac{0.5 - \hat{p}}{\sqrt{(n+1)\hat{p}(1-\hat{p})/(n-1)}}\right), \quad (3) \end{aligned}$$

where $F_{t(n-1)}(\cdot)$ is the $t_{(n-1)}$ cumulative distribution function. Then, $E[\mathcal{Y}_i^*]$ under rounding (RND) is

$$1 - E\left[F_{t(n-1)}\left(\frac{0.5 - \hat{p}}{\sqrt{(n+1)\hat{p}(1-\hat{p})/(n-1)}}\right)\right].$$

Because $n\hat{p} = Y \sim \text{Bin}(n, p)$, we have

$$\begin{aligned} E[\mathcal{Y}_i^*] &= 1 - \sum_{y=0}^n \binom{n}{y} p^y (1-p)^{n-y} F_{t(n-1)} \\ &\quad \times \left(\frac{0.5 - y/n}{\sqrt{(n+1)y(n-y)/[n^2(n-1)]}}\right). \end{aligned}$$

Although this expression is not simple, it can easily be calculated on a computer, without needing any numerical approximations.

Looking at (3), for $E[\mathcal{Y}_i^*] = p$, we must have, under the FN imputation model,

$$E_{\hat{p}}\left[1 - F_{t(n-1)}\left(\frac{0.5 - \hat{p}}{\sqrt{(n+1)\hat{p}(1-\hat{p})/(n-1)}}\right)\right] = p. \quad (4)$$

A sufficient condition for (4) to be true is

$$1 - F_{t(n-1)}\left(\frac{0.5 - \hat{p}}{\sqrt{(n+1)\hat{p}(1-\hat{p})/(n-1)}}\right) = \hat{p}, \quad (5)$$

When $\hat{p} = 0.5$, Equation (5) is true. As \hat{p} deviates further from 0.5, the bias $E[\mathcal{Y}_i^* - p]$ tends to get larger.

We have found that RND leads to the most bias in the tails (small or large p). Figure 1 displays the bias ($E[\mathcal{Y}_i^* - p]$) for values of the true p from 0.001 to 0.50, with $n = 1,000$. The relative bias of $E[\mathcal{Y}_i^*]$ can be calculated using

$$\text{RB}(p) = 100 \left(\frac{E[\mathcal{Y}_i^*] - p}{p}\right).$$

The relative bias in $E[\mathcal{Y}_i^*]$ can be large when p is small (see Figure 2). For example, when $p = 0.05$, $E[\mathcal{Y}_i^*]$ is more than 50% too small. As p gets closer to 0.5, the relative bias goes to zero, although it takes a little bit of an upturn when p is between 0.16 and 0.30. We note that the FN estimator $E[Y_i^*]$ is unbiased, and is not the cause of the bias in $E[\mathcal{Y}_i^*]$.

The magnitude of the bias in the estimate of p using the RND method also depends on the fraction of missing observations:

$$E[\hat{p}^*] = \left(\frac{n}{N}\right)p + \left(1 - \frac{n}{N}\right)E[\mathcal{Y}_i^*].$$

Figure 3 displays the expected value of \hat{p}^* for a variety of missing data proportions and true values of p (0.01, 0.05, and 0.08), for $N = 1,000$. With a high fraction of missingness, rounding can lead to relatively large biases. For example, if one was interested estimating the probability of stomach ulcers from aspirin, which may be around 5%, and 40% of the data were missing, then one could seriously underestimate the probability using rounding. This, in turn, would lead to a large underestimation in the number of people with ulcers from aspirin. For every 100,000 subjects taking aspirin, the number of ulcers would be underestimated by 1,189.

3. DISCUSSION

We have described a simple setting, where it is feasible to correctly specify an imputation model for a missing dichot-

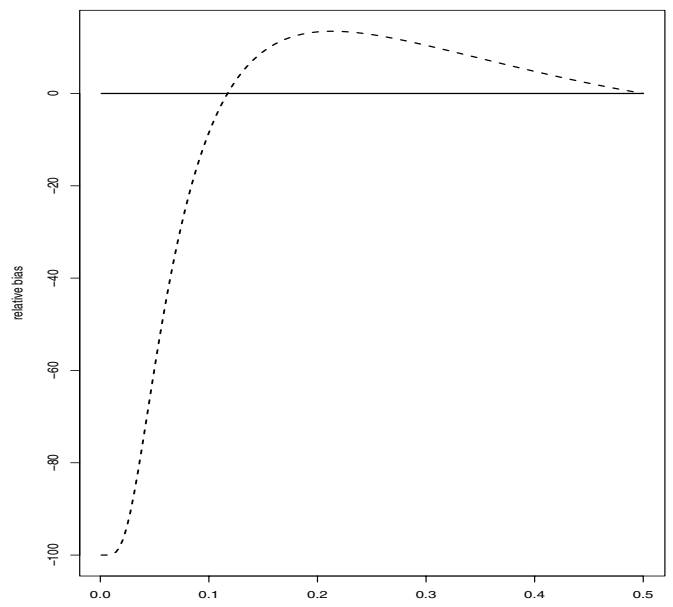


Figure 2. Relative bias ($100*(E[\mathcal{Y}_i^*] - p)/p$) of RND imputation for various probabilities.

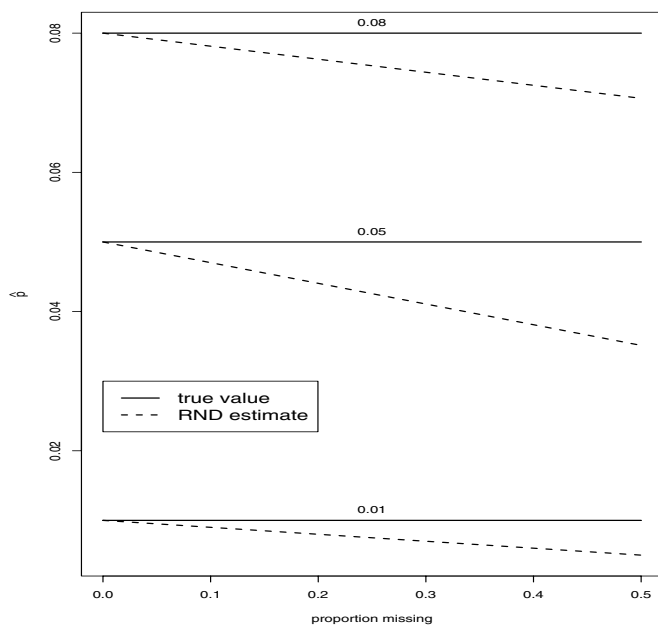


Figure 3. Expected value of RND imputation for a variety of fraction missing (true values 0.01, 0.05, and 0.08; $N = 1,000$).

omous variable. For example, a model could be posited for the probability that $Y_i = 1$ (for $i > n$). For each imputation, a uniform (0,1) random variable could be sampled, and $\mathcal{Y}_i = 1$ if the uniform random variable is greater than the estimated probability, and 0 otherwise, as described by Rubin (1987, p. 169, example 5.3). This type of model is supported by Solas (using discriminant multiple imputation) and the S-Plus 6.0 missing data library (using the log-linear or conditional Gaussian models), and should be used when possible. If one uses a discrete model for imputation, then not only will the bias be minimized, but the results from the multiple imputations will produce correct inferential statements (e.g., confidence intervals and p values).

However, in more complicated settings, with multiple variables with missing values, some of which are continuous, and others categorical, correct specification of the imputation model is difficult. Such mis-specification can lead to bias. The lack of flexible models for the joint distribution of continuous and categorical variables may encourage analysts to adopt methods such as rounding along with an assumption of multivariate normality.

Although this note considered a simpler model, it is plausible that use of rounding in a more complicated setting may also yield nontrivial bias.

This article shows that rounding to make an imputed value “plausible” can actually cause more bias than using the original “implausible” imputation value. For simple models, it is straightforward to correctly specify the imputation distribution to avoid this problem, but in more realistic settings, this requires a complicated joint distribution of continuous and discrete variables. While Schafer (1997, chap. 6) summarized evidence that slight departures from normality tend to yield robust inferences, this evidence was not definitive, and we do not recommend the use of the FN method or the RND method for discrete data.

We would expect “rounding” to cause much less bias for discrete outcomes with many levels, since the effect of rounding relative to the range of the levels is small, which may reduce the relative bias. However, the range of possible missing data configurations and distributions is extremely wide, and the bias will depend on the given configuration. This note shows that rounding should not be used indiscriminately. If possible, we recommend imputing a discrete variable directly from a discrete distribution, such as those available in Solas and S-Plus.

[Received September 2001. Revised July 2003.]

REFERENCES

- Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.
- Rubin, D. B. (1978), “Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse,” in *Proceedings of the International Statistical Institute*, Manila, pp. 517–532.
- (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- 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.
- SAS Institute (2001), *SAS/Stat User’s Guide, Version 8.2*, Cary, NC: Author.
- Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.
- Schimert, J., Schafer J. L., Hesterberg, T., Fraley, C., and Clarkson, D. B. (2000), *Analyzing Data with Missing Values in S-Plus*, Seattle, WA: Data Analysis Products Division, Insightful Corp.
- Sinharay, S., Stern, H. S., and Russell, D. (2001), “The Use of Multiple Imputation for the Analysis of Missing Data,” *Psychological Methods*, 6, 317–329.
- SOLAS, Statistical Solutions, Inc. (2001), Cork, Ireland.