Journal of Official Statistics, Vol. 23, No. 4, 2007, pp. 485–491

Rejoinder

Jan F. Bjørnstad

I would like to thank the discussants for many insightful and constructive comments, definitely enlarging the scope of my article. Especially, I find the many suggestions for future research relevant and interesting. In this rejoinder I will concern myself only with comments about multiple imputation (MI) methods as such. Some comments deal with situations where there is doubt about the appropriateness of using MI at all in estimating variances and constructing confidence intervals. This can obviously happen, but since the article is only concerned with cases where imputation is used to handle nonresponse and repeated imputations are feasible and can be applied, I shall only discuss such cases in this rejoinder.

Skinner describes in a very clear way the problems with different approaches using MI in official statistics, referring to several interesting additional articles. Skinner points out the important fact that using Rubin's combination formula may lead to biased variance estimation even with Bayesian MI for some estimation problems that may be of interest in official statistics, for example domain estimation as shown by Kim, Brick, Fuller, and Kalton (2006). This underlines even more so the need for adjusting Rubin's formula with regard to the specific estimation problem at hand.

One important comment by Skinner concerns Section 6 and the Theorem. As mentioned earlier in Section 6, I am only considering the MCAR case and hot-deck imputation, which comprise the simplest situation in order to categorize when it is possible to use k as the inverse of the response rate. I should have stated these two assumptions in the Theorem as well to avoid any misunderstanding. I regard this result more as an illustration of how to approach the problem mentioned in the title of Section 6. As mentioned by Skinner, a number of dimensions of generality would certainly be useful to research further regarding this issue. In the last paragraph Skinner points the way to future possible research and makes several important points with which I agree.

Skinner mentions the need to compare this non-Bayesian MI method to alternative methods for variance estimation with imputed data. I think it is especially interesting to make a comparison with resampling methods for single imputation since that is the typical case in many NSIs. Münnich discusses the practical issue of computational burden, important in large-scale surveys, and shows that it is possible to reduce the computational burden by a significant amount by using the non-Bayesian MI method instead of more direct bootstrap methods for single imputation.

An essential point discussed by Münnich concerns the sensitivity of the variance inflation constant k, whether one can apply the MI method to a set of estimators simultaneously based on one overall constant for a given problem and m data sets.

This problem is addressed in Section 6 in the simplest case. The simulation study by Münnich to address this problem in a realistic survey, using two hot-deck methods, is very interesting. Münnich is able to show that, with a calibration estimator, it is indeed possible to use $k = 1/(1 - f)$ for more complicated cases and estimators than considered in Section 6. In the survey presented by Münnich $1/(1 - f) = 1.33$ while the k-value from Formula (5), k_{opt}, for the six different situations, lies between 1.25 and 1.43. In four cases k_{opt} is less than 1.33. In fact, since the ratio k_{opt}/k lies in the interval (0.94, 1.07) for all six cases, one would expect that with the simple $1/(1 - f)$, the MI method would give quite a good approximation of the standard error for the calibration estimator. The two cases, where k_{opt} is larger than 1.33, deals with a more heterogeneous population than in the other four cases. This is to be expected, since using $1/(1 - f)$ as a measure of missing information is better suited for homogeneous populations.

The study by Münnich is encouraging for the non-Bayesian MI approach and it may seem that in many complicated situations one can use $k = 1/(1 - f)$ when using a "good" estimator. On the other hand, using an estimator like the Horvitz-Thompson estimator, Münnich shows that this simple value of k is inappropriate. So the issue should certainly be studied further.

An important point made by Chambers is that the variance formula (1) is variable specific, in that k will depend on the response rate for the specific variable or something similar. This cannot be avoided with non-Bayesian imputations. So a secondary analyst needs this variable specific information when imputations are nonproper. Chambers also mentions the fact that NSIs mostly use single imputation strategies and one has to work out the actual variance and variance estimate. This is, of course, correct and the point of combining several imputed data sets by standard analyses is exactly that, there is then no need to work out the actual variance for the imputed estimator. One has a simple "bootstrap" type method of variance estimation. So the recommendation to NSIs should be to use MI when feasible in order to do valid statistical analysis.

I agree with Chambers's contention that most nonresponse mechanisms are MAR (at best) and that there is a need for a general development of this case, even though Section 5 deals with four specific MAR cases.

Laaksonen gives a good description of the current state of affairs at NSIs with respect to MI and the problem of finding good single imputation methods. As he correctly points out, the difficult and most important task is to create imputations that reduce nonresponse bias and display variability close to the true one, before attempting to repeat the imputations. Laaksonen mentions that one reason that MI is not used in NSIs is that it is difficult to find any good imputation method for many practical situations. This is, of course, true regarding unit nonresponse which is typically handled by weighting approaches like poststratification and calibration. For item nonresponse, however, I do believe that most NSIs use some sort of imputation even though it may be a rather simple method like stratified hot-deck or nearest neighbour imputation. Then non-Bayesian MI gives us a way to estimate the variance that includes uncertainty due to imputation.

An important issue raised by Laaksonen is how to make imputations in such a way that k is easily determined. For example, for the simple k being the inverse of the response rate, what kind of imputations would be "correct?" The Theorem in Section 6 says that hot-deck typically is OK in the case of nonresponse being MCAR, but we also have the same k for the residual hot-deck imputation under the ratio model in Section 3.2. So the result in Section 6 can be generalized. As also mentioned by Skinner and Münnich, this type of problem should be an important task for future research.

Laaksonen presents two interesting examples. The case with the continuous variable seems to show, if the imputation-based estimates and reweighted estimates are comparable, that for the type of nearest neighbour imputation methods used here, k should be larger than the inverse of the response rate.

The second example concerns binary variables with nonignorable response mechanism. It is a case of follow-up sampling with no nonresponse at the second phase and can be described as follows.

A simple random sample of size n from a finite population of size N is assumed. The variable of interest y is binary, 0 or 1, with θ being the "success" rate in the population, $\theta = \sum_{i=1}^{N} y_i/N$. Nonignorable response mechanism is assumed; the response indicators R_i are independent and the probability of response is assumed to depend on the y-value. Let $p_1 = P(R_i = 1|y_i = 1)$ and $p_0 = P(R_i = 1|y_i = 0)$. The response sample is s_r with size n_r and the observed sample mean is $\bar{y}_r$. At the second phase we take a follow-up sample s_f of the nonrespondents; a simple random sample of size n_f. It is assumed that there is no nonresponse in the follow-up sample. Let $\bar{y}_f$ be the sample mean in the follow-up. Now the nonresponse rate is equal to $f = (n - n_r)/n$. The "response rate" in the follow-up is given by $n_f/(n - n_r)$. Let the corresponding planned "nonresponse rate" among the missing observations at the first phase be denoted by $f_{mis} = 1 - n_f/(n - n_r)$. The imputation method is hot-deck from the follow-up sample. That is for each unit i outside the follow-up sample the imputed value y_i^* is drawn at random from respondents of the second phase sample.

Laaksonen lets k be $1/(1 - f)$. However, as will be shown, for this case we in fact have

$$k = \frac{n - n_r}{n_f} = \frac{1}{1 - f_{mis}} \tag{I}$$

This result holds both for the design-based and model-based approach. In the example given by Laaksonen, $f = 1/3$ and $f_{mis} = 0.6$ such that $1/(1 - f) = 1.5$ while the correct value of k is $1/(1 - 0.6) = 2.5$. This means that the estimated variance of $\bar{\theta}^*$ should be about 6,600 with SE = 81.2 compared to Rubin's method which gives an estimated SE-value of 73.5. With $k = 1.5$, SE = 76.2.

Proof of (I), Design Approach

In this case, the basic estimator is $\hat{\theta} = \bar{y}_s = \sum_s y_i/n$ and the imputed estimator is given by

$$\hat{\theta}^* = \{n_r\bar{y}_r + n_f\bar{y}_f + (n - n_r - n_f)\bar{y}^*\}/n$$

where $\bar{y}^* = \sum_{s-s_r-s_f} y_i^*/(n - n_r - n_f)$. Clearly, $E(Y_i^*|y_{obs}) = \bar{y}_f$ and $Var(Y_i^*|y_{obs}) = \bar{y}_f(1 - \bar{y}_f)$. It follows that $E(\hat{\theta}^*|y_{obs}) = \{n_r\bar{y}_r + (n - n_r)\bar{y}_f\}/n$ and $Var(\hat{\theta}^*|y_{obs}) = (n - n_r - n_f)\bar{y}_f(1 - \bar{y}_f)/n^2$.

Conditional on (s,s_r), s_f is a simple random sample from $(s - s_r)$. Hence, with $\bar{y}_{mis} = \sum_{s-s_r} y_i/(n - n_r)$, then

$$E\{(n - n_r)\bar{y}_f|s, s_r\} = (n - n_r)\bar{y}_{mis} \text{ and}$$

$$Var(\bar{y}_f|s, s_r) = \frac{\bar{y}_{mis}(1 - \bar{y}_{mis})}{n_f} \cdot \frac{n - n_r - n_f}{n - n_r - 1} \approx \frac{\bar{y}_{mis}(1 - \bar{y}_{mis})}{n_f} \cdot f_{mis}$$

We see that $\hat{\theta}^*$ is unbiased: $E(\hat{\theta}^*) = n^{-1}EE\{n_r\bar{y}_r + (n - n_r)\bar{y}_f|s, s_r\} = E(\bar{y}_s) = \theta.$

This also means that (3) is satisfied, since $\hat{V}^* = \hat{\theta}^*(1 - \hat{\theta}^*)(1 - \frac{n}{N})/n$ and $Var(\hat{\theta}) = \theta(1 - \theta)(1 - \frac{n}{N})/n$. The terms in Formula (5) needed to determine k are now

(i) $VarE(\hat{\theta}^*|Y_{obs}) = n^{-2}[VarE\{n_r\bar{y}_r + (n - n_r)\bar{y}_f|s, s_r\} + EVar\{n_r\bar{y}_r$
$+(n - n_r)\bar{y}_f|s, s_r\}] \approx Var(\hat{\theta}) + n^{-2}E\{(n - n_r)\bar{y}_{mis}(1 - \bar{y}_{mis})f_{mis}/(1 - f_{mis})\}$

(ii) $EVar(\hat{\theta}^*|Y_{obs}) = n^{-2}E\{(n - n_r)f_{mis}\bar{y}_f(1 - \bar{y}_f)\}$
$= n^{-2}EE\{(n - n_r)f_{mis}\bar{y}_f$
$(1 - \bar{y}_f)|s, s_r\} \approx n^{-2}E\{(n - n_r)\bar{y}_{mis}(1 - \bar{y}_{mis})f_{mis}\}$

It follows from Formula (5) that we can use $k = 1/(1 - f_{mis})$.

Proof of (I), Model-based Approach

Model assumption: $\theta = P(Y_i = 1)$ and the Y_i's are independent. In the model-based approach the statistical analysis is conditional on s. Now $\theta = E(T/N)$, with $T = \sum_{i=1}^{N} Y_i$, is to be estimated. Now, $Var(\hat{\theta}) = \theta(1 - \theta)/n$. $\hat{V} = \hat{\theta}(1 - \hat{\theta})/n$ and $\hat{V}^* = \hat{\theta}^*(1 - \hat{\theta}^*)/n$.

$$E(\hat{\theta}^*) = n^{-1}\left[\sum_s E(R_iY_i) + E\{(n - n_r)\bar{Y}_f\}\right] = n^{-1}\left[n\theta p_1 + E\{(n - n_r)E(\bar{Y}_f|n_r)\}\right]$$

Let $\pi = P(R_i = 1) = \theta p_1 + (1 - \theta)p_0$. Then the R_i's are Bernoulli variables (n, π). This implies that

$P(R_i = 1|n_r) = n_r/n$. Let I_i be the indicator for the second phase sample s_f. Then

$$E(\bar{Y}_f|n_r) = n_f^{-1}E\left(\sum_s I_i(1 - R_i)Y_i|n_r\right) = n_f^{-1}nP(I_i = 1|n_r)P(R_i = 0|n_r)P(Y_i = 1|R_i = 0)$$

Let $P(Y_i = 1|R_i = 0) = \theta(1 - p_1)/(1 - \pi) = \theta_0$. Then we see that

$$E(\bar{Y}_f|n_r) = n_f^{-1}n\frac{n_f}{n - n_r} \cdot \frac{n - n_r}{n}\theta_0 = \theta_0 \Rightarrow E\{(n - n_r)E(\bar{Y}_f|n_r)\} = \theta_0(n - n\pi) = n\theta(1 - p_1)$$

and $E(\hat{\theta}^*) = \theta$. This also means that (3) holds. It is readily seen that

$$EVar(\hat{\theta}^*|Y_{obs}) = n^{-2}E\{(n - n_r - n_f)\bar{Y}_f(1 - \bar{Y}_f)\}$$
$$= n^{-2}EE\{(n - n_r)f_{mis}\bar{Y}_f(1 - \bar{Y}_f)|n_r\} \approx n^{-2}\theta_0(1 - \theta_0)E\{(n - n_r)f_{mis}\}$$

It can be shown that $Cov(\bar{Y}_r, \bar{Y}_f | n_r) = Cov(\bar{Y}_r, \bar{Y}_{mis} | n_r) = 0$. Moreover, $E(\bar{Y}_f | n_r) = E(\bar{Y}_{mis} | n_r) = \theta_0$. Then

$$VarE(\hat{\theta}^* | Y_{obs}) - Var(\bar{Y}_s) = n^{-2} E\left[(n - n_r)^2 \{Var(\bar{Y}_f | n_r) - Var(\bar{Y}_{mis} | n_r)\}\right]$$

Now, it is easily shown that $Var(\bar{Y}_f | n_r) = \theta_0(1 - \theta_0)/n_f$ and $Var(\bar{Y}_{mis} | n_r) = \theta_0(1 - \theta_0)/ \, (n - n_r)$.

It follows that $VarE(\hat{\theta}^* | Y_{obs}) - Var(\bar{Y}_s) = n^{-2} \theta_0(1 - \theta_0) E\left[(n - n_r) f_{mis}/(1 - f_{mis})\right]$. From Formula (5) we find that

$$E(k) \approx \frac{n^{-2} \theta_0(1 - \theta_0) E\{(n - n_r) f_{mis}/(1 - f_{mis})\}}{n^{-2} \theta_0(1 - \theta_0) E\{(n - n_r) f_{mis}\}} \approx E\{1/(1 - f_{mis})\} \qquad \text{q.e.d. for (I)}$$

A major part of Thorburn's discussion compares the suggested MI approach to Rubin's MI method. The main purpose of my article is to give an alternative approach when Rubin's method is *not* applicable. I think I make that very clear in Section 1 of the article. In this regard we note, as mentioned by Skinner, that even with Bayesian imputations Rubin's combination formula may lead to biased variance estimation. Therefore, comparisons with Rubin's method are not very relevant when it comes to evaluating the suggested approach in this article. As mentioned by other discussants, it is, however, important in future research to assess the applicability and performance of this non-Bayesian MI approach and make comparisons with other non-Bayesian variance estimation approaches.

The article deals with situations where imputations are the preferred approach to handling nonresponse. Typically, unit nonresponse is treated with weighting methods while imputations are used for item nonresponse. Therefore, it is a misunderstanding when Thorburn claims that the article only considers unit nonresponse. In fact, the article deals mainly with item nonresponse in the sense that unit nonresponse would require simultaneous imputations for all variables in the survey.

In official statistics, imputation is typically done to fill in the missing holes in the data matrix of respondents. Some kind of stratified hot-deck or neighbour imputation are the most common imputation methods, implicitly assuming an MAR response mechanism. Weighting procedures are not really alternative approaches to imputation, but may instead be the full sample estimator. For example, a post-stratified estimator can be a basis estimator $\hat{\theta}$ for the imputation-based estimator $\hat{\theta}^*$, see e.g., Belsby, Bjørnstad, and Zhang (2005).

Thorburn makes several interesting points about Rubin's MI method in his Section 3. However, it is not correct that in hypothesis testing you get the correct p-value by taking the average of the m p-values. Even with Bayesian MI each complete sample does not display sufficient variation, resulting in p-values that are too small. See e.g., Li, Raghunathan, and Rubin (1991) and Pedlow and Meng (1994). The same happens, of course, in non-Bayesian MI. One simply has to do one hypothesis test on the m combined data sets using (5) to determine the correct standard error.

Thorburn discusses in his Section 4.1 the same type of sampling and nonresponse mechanism as Laaksonen. So Formula (I) in this rejoinder is here the correct value of k in the case of binary variables. Note that (I) holds also for continuous variables, provided (3) holds.

Regarding the issue in Section 6 of the article, for the two estimation problems considered by Thorburn in Section 4.4 we note that $k = 1/(1 - f)$ for both estimands in the case of hot-deck imputation. The value $1/(1 - f_X)$ for the first estimation problem is when we use residual hot-deck imputation. This notwithstanding, the point made is important. The value of k may differ for different estimands and, as mentioned earlier, this issue should be studied further.

In Section 5.3 of his discussion, Thorburn observes what the covariance matrix of the imputed values needs to be in order to use Rubin's combination formula. This can be achieved by drawing from Bayes posterior or using methods like the MV method suggested by Rubin and Schenker (1986). Of course, a regular hot-deck method does not achieve this. However, also the suggestion by Thorburn does not work when it comes to achieving the desired covariance matrix. In fact, it does not make any sense since it requires that the response sample and nonresponse sample be of the same size and even then it does not achieve the desired covariance matrix. However, the following imputations will work, letting $\bar{Y}'_i$ be the hot-deck values from Example 3.1,

$$Y_i^* = \sqrt{n_r/(n_r - 1)}\{Y'_i + \bar{y}_r - (1 + \sqrt{n/n_r})\bar{Y}'\} \text{ where } \bar{Y}' = \sum_{s-s_r} Y'_i/(n - n_r)$$

Mass imputation is mentioned as a possible area for multiple imputation. First of all, I think a better term is mass prediction, since each value outside the sample is given a predicted value. Clearly, this is not feasible in most surveys by NSIs. Moreover, the question really is whether it is desirable to do multiple predictions for units outside the sample. One should keep in mind that the multiple imputations are done to handle uncertainty due to item nonresponse in the sample. Also it may cause complications that the conditional distribution for the values outside the sample, given the observed data, is typically not the same as for the missing data in the sample.

In order to develop a theory for non-Bayesian MI it is natural to start by studying the problem with simple response mechanisms such as MCAR and MAR even though, as Thorburn mentions, valid statistical analyses can be performed by ignoring the missing values in these cases. In fact, the MI method for such simple cases may serve as a valid approximate method for more complicated situations as shown in the Comment by Münnich. It is also important to note that imputation is common in survey practice –: for example in surveys done by NSIs imputation is typically done to achieve a complete units-by-variables data matrix for all survey variables. Hence, one needs to develop statistical methods that take into consideration imputation uncertainty whether or not the assumed response mechanism is ignorable. The analyst can then, with a valid MI procedure, combine standard statistical analyses on the complete data sets. Being able to use standard statistical analysis was historically one important reason for imputing for missing data (see e.g., Scheuren 2005), and achieving valid inference with *standard* methods was, of course, also the main reason for developing the Bayesian MI method.

Thorburn draws attention to several interesting areas not discussed in the article, like multivariate imputations preserving covariance structure and imputations for item nonresponse using other survey variables that also may be missing, as auxiliary variables. These are certainly important cases for future studies.

References

Belsby, L., Bjørnstad, J.F., and Zhang, L.-C. (2005). Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey. Survey Methodology, 31, 197–212.

Kim, J.K., Brick, J.M., Fuller, W.A., and Kalton, G. (2006). On the Bias of the Multiple-imputation Variance Estimator in Survey Sampling. Journal of the Royal Statistical Society, Series B, 68, 509–521.

Li, K.H., Raghunathan, T.E., and Rubin, D.B. (1991). 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.

Pedlow, S. and Meng, X.L. (1994). Exploring Hypothesis-testing Procedures with Multiply-imputed Data under Unequal Fractions of Missing Information. Proceedings of the American Statistical Association, Survey Research Methods Section, 595–599.

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.

Scheuren, F. (2005). Multiple Imputation: How It Began and Continues. The American Statistician, 59, 315–319.

Received September 2007