A PARADOX OF MULTIPLE IMPUTATION

Phillip S. Kott, National Agricultural Statistics Service
Room 305, 3251 Old Lee Highway, Fairfax, VA 22030

Key Words: Design, Domain, Model, Repeated imputation, Quasi-random response mechanism

1. INTRODUCTION

Repeated imputation as championed by Rubin (1987) provides a method of adjusting for survey nonresponse. When done correctly, inferences drawn from repeatedly imputed data sets are statistically valid under the right conditions. Some recent attempts to question the validity of repeated-imputation inferences (e.g., Fay 1992) have not fully understood what those conditions are.

A recent article by Meng (1994) tries to clarify the issues surrounding Rubin's repeated-imputation strategy. It fails, however, to reveal the following little understood paradox which occurs when the methodology is applied to weighted survey data: although the imputations themselves are often based on models of variable behavior, variance estimates derived from repeatedly imputed data sets are *not* conditioned on realized survey respondents as is typical in model-based sampling theory (see, for example, Royall and Cumberland 1981). Rather, variance estimation relies on the assumption of a quasi-random response mechanism.

A simple example is introduced in Section 2 that will illustrate this paradox. Section 3 addresses the properties of repeated-imputation methodology in estimating the variance of a theoretical estimator proposed in the previous section. Section 4 discusses a variant of the example from Section 2 and shows that the survey weights themselves are not the issue. The section It also provides a brief discussion.

A note on terminology is in order before proceeding. The more common term "multiple imputation" is used by Rubin to describe a broader class of imputation methodologies than the repeated imputation techniques under discussion here. As a result, except when referring to the literature, the latter expression ("repeated imputation") will be used exclusively in the text.

2. THE EXAMPLE

Suppose one wants to estimate the mean of a domain using a deeply stratified simple random sample. The domain of interest is the union of several design strata with differing sampling fractions. The domain is viewed as fairly homogeneous, so much so that it constitutes a single group for imputation purposes.

The usual design-based, full sample estimator for the domain mean, T, is $t_F = \sum_S w_i y_i / \sum_S w_i$, where S is the sample within the domain, w_i is the inverse of the selection probability for unit i, and y_i is value of interest for i. Since $\sum_S w_i$ is a constant under stratified simple random sampling, t_F is design unbiased. If we assume a model in which the y_i are uncorrelated random variables with mean μ and variance σ^2, t_F is also model unbiased.

Now suppose there is some unit nonresponse. Let R be the respondent sample in the domain, and M be its complement. Following the spirit of repeated imputation as described in Rubin (1987), we can use the model to fill in the missing y-values in M and create a completed sample. This process can the be repeated any number of times. In particular, we create the v'th completed sample $(v = 1, \ldots, V)$ by replacing each missing y_j with

$$y_{jv}^* = \sum_{i \in R} a_i y_i / \sum_{i \in R} a_i + e_{jv} + e_{Ov} [\sum_{i \in R} a_i^2 / (\sum_{i \in R} a_i)^2]^{1/2}, \quad (1)$$

where the e_{uv} are uncorrelated random variable with mean zero and variance s^2, and s^2 is an unbiased estimator for σ^2. For now, we will let the a_j in equation (1) be arbitrary.

Let the domain mean estimator for the v'th completed sample be $t_{(v)} = \sum_S w_j y_{jv}^* / \sum_S w_j$, where $y_{jv}^{(*)} = y_{jv}^*$ when $j \in M$, and $y_{jv}^{(*)} = y_j$ otherwise. One important theoretical construct is the average value of $t_{(v)}$ across an infinite number of completed samples; that is $t_{(\infty)} = plim_{V \to \infty} (\sum t_{(v)} / V)$. It is not hard to see that

$$t_{(\infty)} = (\sum_{i \in S} w_i)^{-1} [\sum_{i \in R} w_i y_i + \sum_{i \in M} w_i (\sum_{j \in R} a_j y_j / \sum_{j \in R} a_j)].$$

When the y_i are uncorrelated random variables with equal variances, $t_{(\infty)}$ has the least model variance as an estimator for t_F when the a_i are all equal. This suggest we set all the a_i in equation (1) to the same value, say unity.

One of the conditions for the imputation in equation (1) to be what Rubin calls "proper" is that $t_{(\infty)}$ be a nearly (i.e, asymptotically) unbiased estimator for t_F under the assumed *response* mechanism (Rubin 1987, p. 118, equation (4.2.5)). We have not as yet stipulated a response mechanism; the model we have been assuming

involves the y-values of the units, not their probabilities of response. One popular response mechanism posits that every sampled unit in the domain under investigation has an equal probability of response. Under this quasi-random response model, one can easily show that $t_{(\infty)}$ is nearly unbiased as an estimator for t_F when $a_i = w_i$ *but not when the* a_i *are all equal* (recall that in our example the sampling fractions vary across design strata).

When $a_i = w_i$, $t_{(\infty)}$ can be expressed as $\sum_R w_i y_i / \sum_R w_i$. Although developed here for theoretical purposes, this version of $t_{(\infty)}$ has been used in practice. See, for example, Kott (1994).

3. THE VARIANCE OF $t_{(\infty)}$

The model variance of $t_{(\infty)}$ as an estimator for μ is:

$$E_M[(t_{(\infty)} - \mu)^2] = E_M[(B + W)^2]$$

$$= E_M(B^2) + E_M(W^2) + 2E_M(BW),$$

where the subscript M denotes expectation with respect to the model governing the y_i, $B = t_{(\infty)} - t_F$, and $W = t_F - \mu$. Now

$$B = t_{(\infty)} - t_F$$

$$= (\sum_{i\in S} w_i)^{-1} \sum_{i\in M} w_i [(\sum_{j\in R} a_j y_j / \sum_{j\in R} a_j) - y_i]$$

$$= (\sum_{i\in S} w_i)^{-1} \sum_{i\in M} w_i [(\sum_{j\in R} a_j e_j / \sum_{j\in R} a_j) - e_i]$$

where $e_i = y_i - \mu$, and

$$W = t_F - \mu$$

$$= \sum_{i\in S} w_i e_i / \sum_{i\in S} w_i$$

Observe that

$$E_M(B^2) = \sigma^2 (\sum_S w_i)^{-2} [\sum_M w_i^2 + (\sum_M w_i)^2 \sum_R a_j^2 / (\sum_R a_j)^2].$$

This is the so-called "between imputation variance." It can be estimated by:

$$\beta = [\sum_{v=1}^{V} t_{(v)}^2 - (\sum_{v=1}^{V} t_{(v)})^2 / V] / (V - 1),$$

which is essentially equation (3.1.4) in Rubin (1987, p. 76). Observe that it is the presence of the e_{0v} and e_{jv} terms in y_{jv}^* as defined by equation (1) that permits β to be a model unbiased estimator for $E_M(B^2)$. For future use, we define $\beta_{(\infty)}$ as $plim_{V\to\infty}\ \beta$.

The so-called "within imputation variance" is $E_M(W^2) = \sigma^2 \sum_S w_i^2 / (\sum_S w_i)^2$, Let:

$$\omega_{(v)} = \sum_{h=1}^{H} [n_h/(n_h - 1)][\sum_{i\in S_h} (w_i y_{iv}^{(*)})^2 - (\sum_{i\in S_h} w_i y_{iv}^{(*)})^2 / n_h], \quad (2)$$

where S_h is the set of n_h sampled units in the domain from stratum h. If there were no nonresponse and the sampling fractions were small enough to ignore, then equation (2) would be the usual unbiased estimator for the *design* variance of t_F. That is its intellectual origin.

Under certain conditions (e.g., when r, the size of R, is large and $\sum_R a_j^2 / [\sum_R a_j]^2$ is small), $\omega_{(v)}$ in equation (2) can be shown to be a nearly unbiased estimator for $E_M(W)^2$. Formally, the model bias of $\omega_{(v)}$ converges to zero as r get arbitrarily large if $\{\max\ ra_i\} / \sum_R a_j$ is bounded. Since there is a different $\omega_{(v)}$ for every v, a reasonable estimator for would be $E_M(W^2)$

$$\omega = \sum_{v=1}^{V} \omega_{(v)} / V.$$

This is equation (3.1.3) in Rubin (1987, p. 76). For future use, we define $\omega_{(\infty)}$ as $plim_{v\to\infty}\ \omega$.

Casual reading of the multiple imputation literature seems to suggest that $\beta + \omega$ is a nearly model unbiased estimator for $t_{(\infty)}$ (see Rubin 1987, p.76, equation (3.1.5)). That would be true in our example only if $E_M(BM)$ is 0. Observe, however, that

$$E(BW) = \sigma^2 (\sum_{i\in S} w_i)^{-2} \sum_{i\in M} w_i [(\sum_{j\in R} a_j w_j / \sum_{j\in R} a_j) - w_i], \quad (3)$$

which will usually *not* be zero given a particular respondent sample within the domain. In fact,

$E_M(BW) >(<)\ 0$ *when*:

$$\sum_R a_j w_j / \sum_R a_j >(<) \sum_M w_j^2 / \sum_M w_j,$$

When $a_i = 1$, this relationship becomes

$$\sum_R w_j / r >(<) \sum_M w_j^2 / \sum_M w_j,$$

where r is the size of R. When $a_i = w_i$, it becomes

$$\sum_R w_j^2 / \sum_R w_j >(<) \sum_M w_j^2 / \sum_M w_j,$$

If we assume a response mechanism in which every sampled unit in the population has an equal response probability, then the right hand side of equation (3) would have an expectation of nearly zero when $a_i = w_i$ under certain conditions (e.g., when r is large). This is because:

$$E_R[\sum_R w_j^2/\sum_R w_j] \approx \sum_S w_j^2/\sum_S w_j \approx E_R[\sum_M w_j^2/\sum_M w_j],$$

where the subscript R denotes expectation with respect to the response mechanism. Under those same conditions, the right hand side of equation (3) would have negative expectation when $a_i = 1$ because:

$$E_R[\sum_R w_j/r] \approx \sum_S w_j/n < \sum_S w_j^2/\sum_S w_j \approx E_R[\sum_M w_j^2/\sum_M w_j],$$

where n is the size of S (since the w_i are not all equal, $\sum_S w_j^2$ must exceed $(\sum_S w_j)^2/n$).

Thus, it appears that repeated-imputation inference is only statistically valid, in the sense of producing a nearly unbiased variance estimator for $t_{(\infty)}$, when the model governing the y_i is combined with a quasi-random response model and then only when the $a_i = w_i$. A saving grace of the repeated imputation variance estimator for $t_{(\infty)}$ when the a_i are all equal is that it is conservative under the twin assumptions of the quasi-random response mechanism and the model governing the y_i.

A more careful reading of Chapter 4 of Rubin (1987) reveals that the "randomization validity" of repeated imputation is the only validity claimed for weighted survey data. By contrast, the Bayesian analysis discussed in Rubin's Chapter 3 would not incorporate weights when centering the posterior distribution of μ given a completed sample.

For repeated imputation methods to be valid in our example, the inference governing a completed sample v can be either model-based or design-based. In our context, $\omega_{(\infty)}$ is required to be a nearly unbiased estimator for either $E_M[(t_F - \mu)^2]$ or $E_D[(t_F - T)^2]$, where D denotes expectation with respect to the original sample design (Rubin's equation (4.2.8), p. 119, is even more limiting). The inference from the respondent sample to the completed sample, however, must be design-based with survey response treated as a second phase of random sampling. In our context, $\beta_{(\infty)}$ must be a nearly unbiased estimator for $E_R[(t_{(\infty)} - t_F)^2]$ (see Rubin's equation (4.2.6), p. 118).

Now the design expectation of $\beta_{(\infty)}$ is a nearly $E_R[t_{(\infty)} - t_F)^2]$ when each sampled unit is equally likely to respond under certain condition, and the model expectation of $\omega_{(\infty)}$ is nearly, $E_M[t_F - \mu)^2]$ but the design expectation of $\omega_{(\infty)}$ under the response model need not be nearly $E_D[(t_F - T)^2]$. To see this last point, consider the extreme case where the y_i are constant within design strata and vary across strata. If there positive fraction of units do not respond, then $\omega_{(\infty)}$ will be positive while $E_D[(t_F - T)^2]$ is zero.

As a result of the properties of β and ω discussed above, the repeated imputation variance estimator for $t_{(\infty)}$, $v_{RI}(t_{(\infty)}) = \beta + \omega$, is nearly unbiased given our assumption about the model governing the y_i and the quasi-random response mechanism, but it is not given the response model and the original sampling design. Moreover, it is not nearly unbiased given the original sample and the model governing the y_i -- an inferential possibility discussed in Rao (1993) and Kott (1994) but not in Rubin (1987).

4. DISCUSSION

We saw that the design expectation of $\omega_{(\infty)}$ under the response model need not be nearly $E_D[(t_F - T)^2]$. As a result, the imputation procedure laid out in Section 2 can not formerly be called "proper" as Rubin defines the term (see Rubin's equation (4.2.8), p. 119). This says more about the limitation of the randomization properties of repeated imputation than about the appropriateness of the imputations themselves. Nevertheless, for completeness sake, we will now discuss a variant of the example in Section 2 under which the imputations are proper but the paradox uncovered in the text remains.

Consider simple random sampling with an ignorably small sampling fraction. Let the target of estimation be the ratio, $T = \sum_P y_i / \sum_P x_i$, where P denotes the population of interest. In the estimator t_F from Section 2, substitute x_i for w_i and $z_i = y_i/x_i$ for y_i. The rest of the analysis follows as before except that now the full sample estimator, $t_F = \sum_S y_i / \sum_S x_i$ is asymptotically design unbiased rather than strictly design unbiased. In addition, equation (2) changes to

$$\omega_{(v)} = [n(n-1)]^{-1} \sum_{i \in S} x_i^2 (z_{iv}^{(*)} - [\sum_{j \in S} x_j z_{jv}^{(*)} / \sum_{j \in S} x_j])^2. \quad (4)$$

In Section 2, all we required of s^2 is that it be a (model) unbiased estimator for σ^2. In order for the imputations to be proper, s^2 must be specified in a tighter fashion here. Let

$$s^2 = \sum_R (y_i - t_{(\infty)} x_i)^2/(\sum_R x_i^2 - 2[\sum_R x_i^3/\sum_R x_i] + [\sum_R x_i^2]^2/[\sum_R x_i]^2),$$

an admittedly odd formulation. An alternative estimator for σ^2 that is not strictly unbiased but satisfies our purposes is

$$s_A^2 = \sum_R (y_i - t_{(\infty)} x_i)^2/(\sum_R x_i^2) = \sum_R [x_i^2 (z_i - t_{(\infty)})^2]/(\sum_R x_i^2).$$

It is demonstrated the appendix that when $a_i = x_i$,

the *design* expectation of $\omega_{(\infty)}$ under the response model is nearly $E_D[(t_F - T)^2]$ for arbitrarily large r under mild conditions (i.e., when the second and third population moments of the x_i are bounded so that, among other things, the difference between s^2 and s_A^2 shrinks towards zero as r gets larger). As a result, the imputations are proper, and $v_{RI}(t_{(\infty)}) = \beta + \omega$ is a nearly unbiased estimator for the design variance of $t_{(\infty)}$ under the original sampling design and the response model.

In this new example, as in the original, the fundamental paradox remains: $v_{RI}(t_{(\infty)})$ is *not* a nearly unbiased estimator given a particular respondent sample when expectations are defined with respect to the same model that generated the imputations in the first place. It is of some interest to note that the full sample estimator, t_F, in the new example is not "weighted" in the usual design-based sense of the term. It appears the paradox has more to do with the inefficiency from the model-based perspective of the nearly design unbiased t_F than with survey weights *per se*.

The practical importance of the paradox in a world where all models fail is an open question. The size of $E_M(BW)$ in equation (3), although not strictly near zero, will usually be small compared to the total model variance being estimated. Nevertheless, for univariate statistics based on complex survey data in the presence of nonresponse, recent work on the jackknife (partially reviewed in Rao 1993) shows greater theoretical promise than repeated-imputation inference can hope to deliver.

What remains unrivaled is the ability of repeated imputation techniques to handle multivariate statistics based on complex survey data with complicated patterns of nonresponse. Even if repeated-imputation inferences are not always exact or even near exact in a strict asymptotic sense, they may be good enough for scientific purposes. Certainly, repeated imputation has no serious competitors at the present time.

REFERENCES

Fay, R. E. (1992), "When Are Inferences from Multiple Imputation Valid?" *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 227-232.

Kott, P.S. (1994), "Reweighting and Variance Estimation for the Characteristics of Business Owners Survey," *Journal of Official Statistics*, 407-418.

Meng, X.L. (1994), "Multiple-Imputation Inferences with Uncongenial Sources of Input," *Statistical Science*, 538-558,

Rao, J.N.K, (1993), "Jackknife Variance Estimation With Imputed Survey Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Vol. 1, 31-40.

Royall, R.M. and Cumberland, W.G (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," *Journal of the American Statistical Association*, 66-77.

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

APPPENDIX: Sketch of a proof that when $a_i = x_i$, $\omega_{(\infty)}$ in the *Discussion* is nearly a design unbiased estimator for $E_D[(t_F - T)^2]$ under the assumed response model

Starting with equation (4),

$$\begin{aligned}
\omega_{(v)} &= [n(n-1)]^{-1}\sum_{i\in S} x_i^2(z_{iv}^{(*)} - [\sum_{j\in S} x_j z_{jv}^{(*)}/\sum_{j\in S} x_j])^2 \\
&= [n(n-1)]^{-1}\sum_{i\in S} x_i^2(z_{iv}^{(*)} - t_{(v)})^2 \\
&\approx [n(n-1)^{-1}1\sum_{i\in S} x_i^2(z_{iv}^{(*)} - t_{(\infty)})^2 \ (since\ t_{(\infty)} \approx t_{(v)}) \\
&\approx [n(n-1)]^{-1}\{\sum_{i\in R} x_i^2(z_i - t_{(\infty)})^2 + \sum_{i\in M} x_i^2 e_{iv}^2\}
\end{aligned}$$

$$(since\ e_{0v}[\textstyle\sum_R x_i^2 / (\sum_R x_i)^2]^{1/2} \approx 0)$$

The last near equality implies

$$\begin{aligned}
\omega_{(\infty)} &\approx [n(n-1)]^{-1}\{\sum_{i\in R} x_i^2(z_i - t_{(\infty)})^2 + \sum_{i\in M} x_i^2 s_A^2\} \\
&= n(n-1)]^{-1}\{\sum_{i\in R} x_i^2(z_i - t_{(\infty)})^2 + \sum_{i\in M} x_i^2 \sum_{j\in R} x_j^2(z_j - t_{(\infty)})^2/\sum_{j\in R} x_j^2\} \\
&= [n(n-1)]^{-1}\{\sum_{i\in R} x_i^2(z_i - t_{(\infty)})^2(1 + m/r)\}
\end{aligned}$$

since $\sum_M x_i^2/M \approx \sum_R x_i^2/R$ under the every-unit-is-equally-likely-to-respond response model.

Continuing,

$$\begin{aligned}
\omega_{(\infty)} &= [r(n-1)]^{-1}\sum_{i\in R} x_i^2(z_i - t_{(\infty)})^2 \\
&\approx [n(n-1)]^{-1}\sum_{i\in S} x_i^2(z_i - t_{(\infty)})^2 \\
&\approx [n(n-1)]^{-1}\sum_{i\in S} x_i^2(z_i - t_F)^2 (since\ t_{(\infty)} \approx t_F) \\
&= [n(n-1)]^{-1}\sum_{i\in S} (y_i - x_i t_F)^2,
\end{aligned}$$

which itself is a nearly design unbiased estimator for $E_D[(t_F - T)^2]$.