MULTIPLE IMPUTATIONS IN SAMPLE SURVEYS - A PHENOMENOLOGICAL BAYESIAN APPROACH TO NONRESPONSE

Donald B. Rubin, Educational Testing Service

A general attack on the problem of nonresponse in sample surveys is outlined from the phenomenological Bayesian perspective. The objective is to develop procedures that are useful in practice. The plan is to impute several values for each missing datum, where the imputed values reflect variation within an imputation model and sensitivity to different imputation models. The analysis of a resultant multi-imputed data set is viewed as simulating predictive distributions of desired summary statistics under imputation models. Three tasks are defined that are needed to create the imputations: the imputation task, the estimation task and the modelling task. The imputation task and estimation task are technical in nature. The modelling task requires the development of new tools appropriate for relating nonrespondents and respondents.

1. Introduction

The problem of nonresponse in sample surveys has recently attracted much interest. Possible reasons for this increased interest include (1) current surveys appear to be suffering more serious problems of nonresponse, (2) there exists a growing awareness that standard methods of handling nonresponse may not be entirely satisfactory, and (3) the statistical issues arising in handling missing data form a fertile area for statistical research both mathematically and computationally.

As reflections of the increased concern with nonresponse, we see as examples recent OMB policy stating that no survey should be approved that anticipates less than a 50% response rate, the formation of the National Academy of Sciences' Panel on Incomplete Data, an increase in applied papers on methods of handling nonresponse, and an increase in mathematical statistical papers on estimation from missing data.

My objective is to develop statistically sound tools for handling nonresponse in general purpose surveys. Hence, I'll be concerned with both theoretical appropriateness and practical utility. This paper will outline my suggestions for a general approach to handling the problems of nonresponse.

Section 2 presents an overview of the ideas and motivates the choice to use several imputed values for each missing datum as a method of simulating predictive distributions of missing values. Section 3 discusses how to analyze a data set with multiple imputed values. Section 4 outlines three tasks needed to create a data set with multiple imputed values: the imputation task imputes values assuming that a model for the data has been chosen and that the posterior distribution of model parameters has been calculated; the estimation task calculates the posterior distribution of parameters assuming a model has been specified; and the modelling task selects a model for the data. The first two tasks are conceptually quite straightforward although the technical details can be somewhat complicated. The last task, choosing appropriate models, requires conceptual development.

2. General Approach to Handling Nonresponse

In order to help motivate my approach to nonresponse, I'll begin by describing a few surveys familiar to me that suffer from nonresponse. Several years ago,ETS conducted a survey of 660 schools for the purpose of studying compensatory reading programs,and needed to obtain the principals' permission to enter the schools the next year for an intensive testing program for the students. Of the 660 principals contacted, by the end of the survey only 472 completed an initial questionnaire indicating willingness to participate. Since the principals knew the purpose of the survey was to study their compensatory reading programs, concern developed that the 188 nonrespondents were systematically different from the 472 respondents, perhaps having students with more severe reading problems.

A second example is a study on cost of caring for terminal cancer patients. In this study, there were several barriers to actually obtaining the costs from the patients. The interviewer first had to obtain permission from the patient's hospital, then from the patient's doctor; then the interviewer was allowed to confront the patient's family and, finally, the patient. In this study there was a very high nonresponse rate, greater than 50%. Again, the suspicion was that the nonrespondents differed systematically from respondents, perhaps being more incapacitated.

Another example of nonresponse is item nonresponse on income questions in the CPS (Current Population Survey). It is certainly possible that those refusing to supply income information systematically differ from those willing to supply it.

What do we mean by handling the problem of nonresponse in examples such as these? I feel that handling nonresponse must mean displaying how different the answers from the surveys might have been if the nonrespondents had responded. Since we cannot know this without obtaining responses from nonrespondents, our objective will be to show how the answers change under a variety of reasonable models. Since "reasonable" is partially determined by individual judgment (and partially by the observed data), effort must be directed not only at showing how answers change under different models, but also at precisely communicating those models that have been used so that the appropriateness of the models can be judged.

2.1 Handling nonresponse by imputation

The general approach to nonresponse (missingness) in surveys that I will take here will be to impute values for missing data (really, several values for each missing datum). The approach that imputes one value for each missing datum is quite standard in practice, although often criticized by some more mathematical statisticians who may prefer to think about estimating parameters under some model.

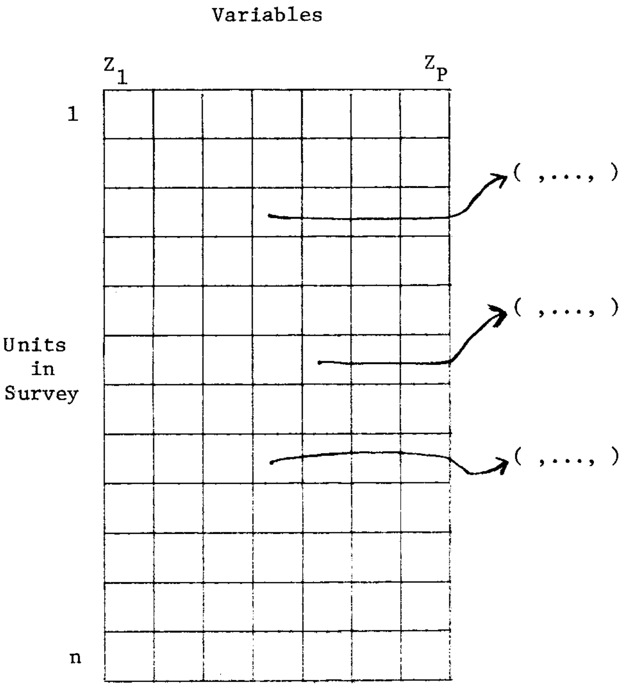
I am sympathetic with the imputation position for two reasons. First, it is phenomenological in that it focuses on observable values. There do not exist parameters except under hypothetical models; there do, however, exist actual observed values and values that would have been observed. Focusing on the estimation of parameters is often not what the applied person wants to do since a hypothetical model is simply a structure that guides him to do sensible things with observed values. Second, in multipurpose surveys, some form of imputation is just about the only practically possible method for handling nonresponse, because the data set will be used to address many questions now and in the future. Remodelling the missing data process each time a new question is to be asked of the data base seems to be impractical, while creating an imputed data set is quite practical.

Of course, (1) imputing *one* value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing); and (2) in order to insert sensible values for missing data we must rely on some model relating unobserved values to observed values. Hence, I see the best practical approach to be one where we can insert more than one value for a missing datum in a way that reflects our uncertainty; the inserted values should reflect variation within a model as well as variation due to a variety of reasonable models. Also, I want to perform the imputation in a formal non-*ad hoc* manner so that I know how to interpret the imputed values; that is, so that I understand precisely the models used and can communicate them precisely to other researchers interested in this data set with its collections of imputed values.

Figure 1 displays the kind of data set I envision forming when there exists nonresponse. Each of M missing data is replaced by a pointer to a vector of length I, giving the imputed values of I different imputations. The first component of each I-vector refers to the first imputation, the second component to the second imputation, and so on. For each of the imputations, there would be a code describing the model/assumptions used for the imputations. The data analyst wanting to study this data set would be obligated to analyze it I times, once for each imputation, and compare the results. More on this in Section 3.

2.2 Our theoretical perspective

Our theoretical approach for performing the imputations will be a phenomenological Bayesian perspective. The foundations of this approach are outlined in Rubin (1978a, 1978b), which provides a natural framework for handling the problem of nonresponse. In the phenomenological Bayesian perspective, the missing values have a distribution given the observed values. Hence, what we really want to impute is not a single value but the predictive distribution of the missing values given the observed values. Such a distribution will of course depend on a model, and displaying the sensitivity of inferences to reasonable choices of models is a key objective.



Model for first imputation = . . .
Model for second imputation = . . .
.
.
Model for Ith imputation = . . .

Figure 1: Data Set with Multiple Imputations for Each Missing Datum

From the phenomenological Bayesian perspective, when mechanisms used to sample units and record data are known (possibly probabilistic) functions of recorded values, the mechanisms are said to be ignorable. The advantage of having ignorable mechanisms is that we can use the usual kinds of models to estimate the unobserved values in the survey; more explicitly, when mechanisms are ignorable, the distribution of the data may be modelled as row exchangeable, that is, essentially independent and identically distributed (i.i.d.), given some model parameters having a distribution. Not surprisingly, in simple problems with ignorable mechanisms, this approach gives the same kinds of answers that classical sampling theory gives; in more complicated problems, I find the phenomenological Bayesian answers more attractive.

When mechanisms are nonignorable (for example, when there exists nonresponse) the phenomenological Bayesian perspective tells us that there is in general a separate model for each group of units with the same pattern of missingness. The trick for drawing inferences when faced with nonignorable mechanisms is to tie the parameters for the different groups of units together so that the values we do see tell us something about the values we do not see. The phenomenological Bayesian perspective tells us that we should try a range of reasonable models, and for each model calculate the predictive distribution of the missing values given the observed values. The idea of imputation is natural in this perspective because by producing several imputed values under a model we simulate the predictive distribution of the missing values under that model.

2.3 Inherent model sensitivity when mechanisms are nonignorable

The values to impute will depend on the models that we decide to use. Nonresponse is a difficult problem because there will be no hard evidence in the data themselves to contradict relevant aspects of a model for nonresponse.

A small example makes this point rather clear. Suppose that we have a population of 1000 units, try to record a variable Z, but half of the units are nonrespondents. For the 500 respondents, the data look exactly half-normal. Our objective is to know the mean of Z for all 1000 units. Now, if we believe that the nonrespondents are just like the respondents except for a completely random mechanism that deleted values (i.e., if we believe that mechanisms are ignorable), the mean of the respondents, that is, the mean of the observed half-normal distribution, is a plausible estimate of the mean for the 1000 units in the population. However, if we believe that the distribution of Z for the 1000 units in the population should look more or less normal, then a more reasonable estimate of the mean for the 1000 units would be the minimum observed value because units with Z values less than the mean refused to respond. Clearly, the data we have observed cannot distinguish between these two models except when coupled with prior assumptions.

Because of this need to rely on prior assumptions, one of the components of a general purpose method of handling nonresponse is the ability to display the sensitivity of answers to a range of models. For example, as we try a variety of reasonable imputation models, we should be able to see whether the variation in answers swamps the usual standard errors that would be associated with the answers. At the other extreme, we should be able to see whether the usual standard errors that are associated with the answers swamp the variability that we see in the answers as we move from one reasonable imputation to another. Performing repeated data analyses on the data set with different imputed values appears to me to be the most natural way to display this sensitivity.

2.4 Survey design considerations

The example in Section 2.3 illustrated the inherent model sensitivity to imputation models when mechanisms are nonignorable. What can be done to reduce such sensitivity? Surveys anticipating nonresponse problems should try to collect background variables recorded for all units that are (1) highly correlated with variables likely to be missing, and (2) related to the reasons for nonresponse (i.e., correlated with missingness indicators).

The first criterion is rather obvious; having recorded variables highly correlated with missing variables implies that it is relatively easy to predict missing values from observed values; i.e., under reasonable models, the predictive distribution of the missing variables given the observed variables has small variance because knowing the values of recorded variables implies almost knowing the values of missing variables.

The second criterion is important for the following reason. If the variables that determine nonresponse are recorded for data analysis, then the probability of the observed pattern of missingness is a function of these recorded values and thus the recording mechanism may be effectively modelled as ignorable. As mentioned earlier, illustrated by the example in Section 2.3, and further explicated in Sections 3 and 4, when mechanisms are ignorable, sensitivity to imputation models is greatly reduced.

Hence, when nonresponse may be a problem, it is a good idea to try to collect background variables that predict missingness and/or predict variables likely to be missing.

3. Data Analysis with Imputed Values

Assume for each of M missing values we have I imputed values, as illustrated in Figure 1. These values are ordered in the sense that we view them as I M-vectors being used to create I completed data sets. This section describes the data analysis strategy we propose for these I completed data sets.

3.1 Analyses for each completed data set

Suppose that there were no missing values. Then there is an analysis or a series of analyses that would have been performed. These analyses may have been aimed at producing summary tables, estimating means and standard errors, performing regressions or factor analyses, and so on. I would usually prefer the analyses to be Bayesian when they are intended to produce inferences to a population from which the current data are considered a sample (actually phenomenological Bayesian). But the issue of proper analysis when there is no nonresponse is not the issue here, so I will not pursue it.

The point I wish to make here is simply that there is some sequence of analyses that would have been performed if there had been no missing data, and in these analyses some summary functions of the data would have been calculated and examined. Call all statistics that would have been calculated the vector $\underline{S}$. Now when faced with missing data and I imputed data sets, the same sequence of analyses should be performed on each of the I completed data sets. That is, we should treat each of the I data sets as if it were the one real data set, and so generate statistics $\underline{S}_1,\ldots,\underline{S}_I$. The variation in answers (in the $\underline{S}_i$) across the I analyses is telling us about the effect of nonresponse on our analyses.

3.2 Variation in answers under a specific imputation model

Suppose first that all I M-vectors of imputed values were generated under the same model for nonresponse. Then the variation in answers across the I data sets reflects variation due to inability, under the model, of the observed data to predict the missing data. More explicitly, the distribution of $\underline{S}_i$, i=1,...,I, simulates the predictive distribution of $\underline{S}$ under that model for nonresponse.

For a trivial example, consider a sample of 100 units, 20 nonrespondents and a binomial Z, where forty of the eighty respondents have Z = 1. Suppose that if there were complete data we would have calculated the proportion of the 100 units with Z = 1. If we chose a model for nonrespondents that asserts Z is i.i.d. binomial ϕ_1 where ϕ_1 has, say, a Beta (1/4,1/4) distribution, then we could take I = 10 draws of ϕ_1 from Beta (1/4,1/4), and for each draw of ϕ_1, we would choose 20 independent and identically distributed binomial observations with probability ϕ_1 of being 1 and calculate S = (40 + sum)/100, where sum is the number of the 20 draws yielding Z = 1.

In some cases, the variation within a model may be small, while in other cases it may be large. When it is large and the imputation model being used is reasonable, sharp inferences will not be possible without further restrictions on the model being used. In such a case, a procedure that produced only one imputation under the model would clearly be leading the data analyst astray. Hence, I feel we must be able to display the variation under an imputation model.

For some models, it may be easy to calculate the predictive distribution of $\underline{S}$ analytically. Usually, I believe that it will be easier to simulate the distribution, especially for general-purpose surveys having $\underline{S}$'s with many components.

3.3 Sensitivity of answers to models

Now suppose that the I filled-in data sets represent K different models used for imputation. Specifically, suppose that $I = \sum_{k=1}^{K} r_k$ where r_k = the number of replications of the kth model, replications in the sense of Section 3.2. We have already discussed that for a fixed model the variation in values of $\underline{S}$ across the r_k replications simulated the posterior distribution of $\underline{S}$ given that model. The variation in these posterior distributions across the K imputation models represents sensitivity to the imputation models. In the example of Section 3.2, an alternative model could be that the prior on ϕ_1 is Beta (1/2,1/4).

If we placed a prior distribution across the K imputation models, then they would really be one model since they would generate one predictive distribution of $\underline{S}$. Often, instead of placing a rather arbitrary prior distribution on the models, it is more enlightening to display the sensitivity to the models. This attack seems especially appropriate in cases like nonresponse having data that cannot directly contradict the models; i.e., the posterior probabilities of the models essentially equal their prior probabilities.

If there exists substantial variation in answers across the K imputation models, then the data analyst who imputed from only one model might be drawing sharp conclusions without realizing that the conclusions are critically dependent on the particular imputation model used. Hence, I feel that we must be able to display sensitivity to reasonable choices of imputation models.

3.4 Base line models assuming ignorable mechanisms

It will usually be wise to include at least one model that assumes ignorable sampling and recording mechanisms; this assumption means that the nonrespondents do not systematically differ from respondents in other than observed ways. Technically, the probability of the observed pattern of missing data and sampled units is known from the observed values. When mechanisms are ignorable, one common data structure may be assumed for all units. Almost all current methods of handling missing data either explicitly or implicitly make this assumption.

In our simple examples of Sections 3.2 and 3.3, if mechanisms are ignorable, ϕ_1 equals ϕ_0, the probability that Z = 1 for respondents. Since the posterior distribution of ϕ_0 with 80 observations depends only modestly on the prior distribution for ϕ_0, there may be no need to use several ignorable models and display the sensitivity to these. In other cases, especially those with multivariate Z, it may be wise to consider several ignorable models (e.g., linear and quadratic regressions, log-linear regressions), to enable an evaluation of sensitivity to prior assumptions about the data structure assumed common across all units.

3.5 Demands on system and user

There are undeniable demands placed on both system and user by requiring I analyses of each completed data set. There must be efficient ways to store the I M-vectors (e.g., by pointers), and the user must be willing to examine the results of the I data analyses for variability of answers within models and sensitivity of answers to different models.

The questions of how many replications of a model are needed and how many different models should be used are difficult if not impossible to answer in general. Perhaps in some cases just a few replications (e.g., five) and one model will be enough to tell the data analyst that nonresponse is a serious problem; for example, suppose that a nonresponse model positing only mild differences between respondents and nonrespondents yields vastly different answers in just five replications. In other cases, five replications and one model may be enough to convince the data analyst that there is no real problem due to nonresponse; for example, suppose that the one model posits potentially quite violent differences between respondents and nonrespondents and yet the five answers differ in only minor ways. Interactive computing may be of great benefit to such analyses.

In all cases, the models used for imputation should be explicitly stated so that results can be unambiguously transmitted between, and evaluated by, researchers. Hopefully, within a particular substantive area of study, experience will suggest which models are most acceptable. Of course, this supposes active study aimed at understanding the relative attributes of respondents and nonrespondents in that research area; e.g., Rosenthal and Rosnow (1975) study the volunteer and nonvolunteer subject in psychological research.

With large data sets and modest computing facilities, it may be necessary to randomly subsample the data set and evaluate the effect of nonresponse on the subsample. In the same way that a randomly drawn sample can tell us about a population, a randomly drawn subsample can tell us about nonresponse problems in the sample.

4. Definition of tasks needed to create imputed values

I see three tasks needed to create the imputed values assumed in Section 3 to exist: the "imputation task", the "estimation task", and the "modelling task". Figure 2 displays how these tasks fit into our general plan for handling non-response. The modelling task chooses a model for the data, the estimation task computes a posterior distribution for the parameters of this model, and the imputation task takes one random draw from the associated predictive distribution of the missing data given the observed data.

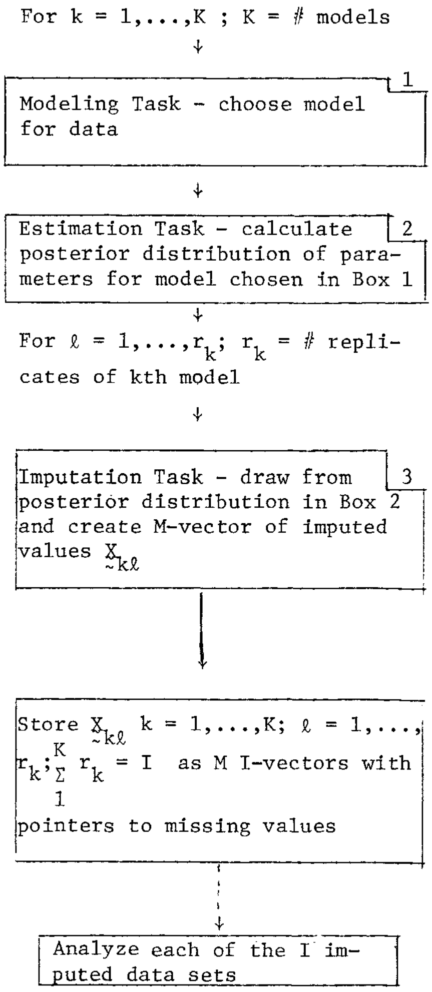

For k = 1,...,K ; K = # models
1
Modeling Task - choose model for data
2
Estimation Task - calculate posterior distribution of parameters for model chosen in Box 1
For ℓ = 1,...,r_k; r_k = # replicates of kth model
3
Imputation Task - draw from posterior distribution in Box 2 and create M-vector of imputed values $X_{k\ell}$
Store $X_{k\ell}$ k = 1,...,K; ℓ = 1,...,r_k; $\sum_1^K r_k$ = I as M I-vectors with pointers to missing values
Analyze each of the I imputed data sets

Figure 2: Creating the Multiple Imputations

4.1 The imputation task

Let Z be the n x p units by variables matrix of data in the survey. The imputation task assumes that (1) Z has been modelled with rows independent (not necessarily identically distributed) given an unknown parameter ϕ and (2) the posterior distribution of the parameter ϕ, pos(ϕ), exists in the sense that a random draw of ϕ from pos(ϕ) can be made. The imputation task then makes one random draw from the associated predictive distribution of missing values given the observed values under that model, and thus creates one M-vector of imputed values to complete the data set. When r replications are desired for a model, the imputation task is performed r times for that model. These replications simulate the predictive distribution of the missing values given the observed values under that particular model.

The imputation task begins by sorting the sampled units by their pattern of missing data. Index these patterns by j=0,.,,,,.J, where j=0 refers to units with no missing data. The phenomenological Bayesian framework tells us that, in general, each pattern of missing data corresponds to a separate i.i.d. model for the data. For the jth pattern of missing data, suppose that each row of Z is modelled as i.i.d. $f(Z|\phi_j)$, j=0,...,J where $\phi_j = \phi_j(\phi)$ and ϕ has posterior distribution pos^j(ϕ).^{j} When mechanisms are ignorable, $\phi_0 = \phi_1 = \ldots = \phi_J$, and thus the rows of Z are not only independent, they are also identically distributed.

For the jth pattern of missing data, partition Z into Z = (V_j,U_j) where V_j are the missing variables and U_j are the observed variables; for j=0, Z=U_0. Since for each unit we must impute values for V_j given the model and the observed values of U_j, we factor the density $f(Z|\phi_j)$ as

$$f(Z|\phi_j) = f(V_j|U_j,\xi_j)\; f(U_j|\eta_k)$$

$$\text{where}\quad \xi_j = q_j(\phi_j)$$

$$\eta_j = \bar{q}_j(\phi_j)$$

and where $q_j(.)$ and $\bar{q}_j(.)$ are the appropriate functions of the parameter ϕ_j corresponding to the partition Z = (V_j,U_j). For example, if Z is normal under $f(\cdot|\cdot)$, then ϕ_j represents the mean and covariance of Z, $\xi_j = q_j(\phi_j)$ represents the regression parameters of V_j on U_j (i.e., the regression coefficients and conditional covariance of V_j given U_j), and $\eta_j = \bar{q}_j(\phi_j)$ represents the marginal parameters of U_j (i.e., the mean and covariance of U_j). If Z is categorical under $f(\cdot|\cdot)$, then ϕ_j represents the cell probabilities defined by Z, ξ_j represents the conditional probabilities of being in cells defined by V_j given membership in the cells defined by U_j, and η_j represents the probabilities of being in cells defined by U_j.

Using this notation, the imputation task is as follows.

A. Draw ϕ_0 from the posterior distribution of ϕ_0, pos(ϕ_0). Call the drawn value ϕ_0^*.

B. For j=1,...,J

(i) draw ϕ_j from pos$(\phi_j|\phi_0 = \phi_0^*, \ldots, \phi_{j-1} = \phi_{j-1}^*)$

(ii) calculate $\xi_j^* = q_j(\phi_j^*)$

Equivalently, letting $\xi_j = q_j(\phi_j)$ and $\eta_j = \bar{q}_j(\phi_j)$,

(i') draw ξ_j from pos$(\xi_j|\phi_0 = \phi_0^*, \xi_1 = \xi_1^*, \ldots, \xi_{j-1} = \xi_{j-1}^*)$

(ii') call the drawn value ξ_j^*

(iii) for each unit with the jth pattern, independently draw the imputed value of V_j from $f(V_j|U_j=U_j^*, \xi_j=\xi_j^*)$ where U_j^* is the unit's value of U_j.

In large surveys with many respondents, pos(ϕ_0) may be nearly point mass; however, the posterior distribution of each ξ_j will generally not be very sharp, even if the jth group is very large, because there exist no data to directly estimate ξ_j. That is, since V_j is entirely missing for the jth group of units and ξ_j is the parameter of the conditional distribution of V_j given U_j, there are no data to estimate ξ_j directly. Hence, even when pos(ϕ_0) is point mass, the values ξ_j^* will vary in replications of the imputation task.

When mechanisms are ignorable, $\xi_j=q_j(\phi_0)$ so that steps B(i) and B(ii) may be replaced by

calculate $\xi_j^* = q_j\ (\phi_0^*)$.

In large surveys with pos(ϕ_0) essentially point mass at ϕ_0^* and with ignorable mechanisms assumed, the values ξ_j^* will not vary in replications of the imputation task, and then variation in the imputed values V_j will be due to inability to predict perfectly V_j from U_j. Hot deck procedures that randomly draw from the matches are essentially drawing from a predictive distribution under a categorical model assuming (1) the posterior distribution of ϕ_0 is point mass (i.e., equals the observed proportions for respondents), and (2) mechanisms are ignorable.

Whether mechanisms are ignorable or not, having a large number of background variables that are recorded for all units and are highly correlated with variables that may be missing reduces the variation in imputed values across repeated imputations. In our notation, the residual variation in the model $f(V_j|U_j,\xi_j)$ used in step B is smaller if U_i includes many variables capable of predicting V_j than if U_j includes only a few relevant variables.

4.2 The estimation task

The estimation task assumes one model, $f(Z|\phi)$ with prior distribution prior(ϕ), has been chosen and computes the posterior distribution of ϕ. Actually, calculating this posterior distribution can be not only analytically intractable but also computationally demanding. Consequently, we often must be satisfied with approximate posterior distributions from which we can easily draw.

Ignorable mechanisms imply that "the missing data are missing at random" and model parameters are "distinct" from missingness parameters (Rubin, 1976). Consequently, when mechanisms are ignorable, the estimation task corresponds to Bayesian computations for posterior distributions of parameters when faced with missing data, ignoring the process that creates missing data. Often special computational programs are needed even with rather simple unrestricted models. We will not delve into this issue because it is not unique to our method of handling nonresponse problems, but arises with parametric inference when faced with missing data. The statistical literature on maximum likelihood/Bayesian estimation of parameters when faced with missing data is now quite extensive. If the respondents greatly outnumber all nonrespondents, little may be lost by estimating ϕ_0 from the complete data alone. Standard hot deck procedures may be thought of as estimating a categorical model assuming mechanisms are ignorable using only the data from the respondents.

When mechanisms are nonignorable, more elaborate, nonstandard models are required and there is essentially no literature on calculating posterior distributions under such special models. Presumably, much statistical/numerical work may be needed to find useful approximations to posterior distributions under such models. The basic problem arises because when mechanisms are nonignorable, each pattern of missingness in general has its own parameter ϕ_j. If the $\phi_0,\ldots,\phi_J$ were not tied together via a prior, only the posterior distributions for ϕ_0 and $\eta_j = \bar{q}_j(\phi_j)$ j = 1,...,J would be modified by the data; the posterior distributions for $\xi_j=q_j(\phi_j)$, j= 1,...,J would then equal their prior distributions unless ξ_j and η_j were dependent _a priori_. Because only the parameters ξ_j are used in the imputation step, the models we wish to use will, via prior restrictions, tie the ϕ_j together in some way and/or tie ξ_j to η_j in some way. These ties may generate nonstandard models and therefore the estimation task for such models may need development. Of course, there are cases of these special models in which the estimation task (or a simple approximation to the estimation task) uses standard tools. Rubin (1977, _JASA_) and Rubin (1977, unpublished) address such situations.

4.3 The modelling task

The modelling task formulates the model $f(Z|\phi)$ prior(ϕ) needed in the imputation task. Assuming ignorable mechanisms, this task is simply a good standard Bayesian formulation of a model appropriate for a multivariate data set. This is not to say the task is easy; it may not be, even with complete data, but the issues that arise are basically the same as arise with complete data.

Although in the context of nonresponse there is an emphasis on obtaining accurate predictions of those missing values that occur in the data matrix, standard tools of mathematical statistics can be applied without much modification. Hence, with ignorable mechanisms, the modelling task does not need new tools.

I must emphasize that in this context "standard" tools do not mean only models commonly used to analyze multivariate data. Rather I mean to include hyperparameter (or Empirical Bayes in the Efron and Morris sense) models that borrow strength by tying parameters together via prior restrictions. Such models play an important role in obtaining good estimates in complete-data problems, and I believe have an even larger role to play in missing data problems, especially when mechanisms are nonignorable.

When mechanisms are nonignorable, however, new tools may be needed. The basic problem arises because, in general, each pattern of missingness has a separate model, and thus we need special models if we want to tie models together. As far as I know, Rubin (1977,JASA) is the only published example of a model that ties respondents to nonrespondents via a nonignorable model.

When building these special models, we should focus on ways to formalize prior knowledge and practical wisdom. I believe that the intelligent use of hyperparameter models is virtually necessary in all but the simplest cases. Here I will simply discuss a few kinds of models in order to indicate the kinds of ties that can be made between respondents and nonrespondents.

When mechanisms are modelled as ignorable, $\phi_0 = \phi_1 = \dots = \phi_J$, and thus $\xi_j = q_j(\phi_0)$ with the strongest possible ties between the parameters for the different patterns of missingness. Consider the simplest case, J=1; see Figure 3. Usually we would model Z so that $\xi_1(\phi_0)$ and $\eta_1(\phi_0)$ are a priori independent and thus estimate $\xi_1(\phi_0)$ only from the respondents. For example, consider the usual linear or log-linear regression models where the distribution of the independent variables is considered "fixed." Note, however, that if ξ_1 and η_1 are dependent a priori, as with some hyperparameter models on Z, the nonrespondent U_1 data help to estimate ξ_1; that is, even though no V_1 data exist for nonrespondents, the nonrespondents' U_1 data are informative about the conditional distribution of V_1 given U_1. With ignorable mechanisms and many respondents, such information will usually be swamped by the stronger information about ξ_1 coming from the respondents, but it may be important when mechanisms are nonignorable.

At the other extreme from assuming ignorable mechanisms, we have $\phi_0, \phi_1, \dots, \phi_J$ a priori independent. Under such a model, we can only learn about $\xi_j = q_j(\phi_j)$ from the observed U_j data for the jth group, that is, $\text{pos}(\xi_j) = \int \text{prior}\ (\xi_j | \eta_j)\ \text{pos}\ (\eta_j)\ d\eta_j$ and $\xi_0, \dots, \xi_J$ are a posteriori independent. Now, prior ties between $q_j(\phi_j)$ and $\eta_j = \bar{q}_j(\phi_j)$ may be quite important if, for example, prior exchangeability arguments could be made among some of the variables. Of course, the data may tell us that $\phi_0, \phi_1, \dots, \phi_J$ should not be modelled as a priori independent; e.g., presumably, an examination of independent estimates will show that they might be more sensibly modelled as arising from a common distribution governed by a hyperparameter.

An example of a simple model with some prior dependence between the $\phi_0, \phi_1, \dots, \phi_J$ makes all ties through ϕ_0; that is,

$$\text{prior}(\phi) = \text{prior}(\phi_0) \prod_{j=1}^{J} \text{prior}(\phi_j | \phi_0).$$

Units	Z Variables	
	U_1	V_1
Respondents j = 0	observed	observed
Nonrespondents j = 1	observed	missing

Figure 3: Simple Case of One Pattern of Nonresponse

If we also add the restriction on $\text{prior}(\phi_j | \phi_0)$ that $\text{prior}(\xi_j | n_j, \phi_0) = \text{prior}(\xi_j | q_j(\phi_0))$, we end up with a model that is easily dealt with in practice when the respondents are a large proportion of the sample. That is, under the model specified for the prior on ϕ, we have for task B(ii'),

$$\text{pos}(\xi_j | \phi_0,\ \xi_1, \dots, \xi_{j-1}) = \text{prior}(\xi_j | q(\phi_0)),$$

and if the respondent sample is large,we have that $\text{pos}(\phi_0)$ may be essentially determined by the respondent data alone. Hence, in this case for usual models, $\text{pos}(\phi_0)$ is relatively easy to calculate and draw from, $q_j(\phi_0)$ is relatively easy to calculate, and $\text{pos}(\xi_j | \phi_0, \xi_1, \dots, \xi_{j-1})$ will be easy to draw from by choosing commonly used prior distributions. The approach is used in Rubin (1977, JASA) when J=1, and suggested in Rubin (1977, unpublished) for $J > 1$.

Of course, just because the model of the previous paragraph is easy to use does not mean it is appropriate. In many cases I feel there will be stronger information in the data than is being reflected in such a model, and we need new hyperparameter models to reflect this information. For example, under exchangeability arguments about variables, if the distribution of U_j appears similar in the jth group and the respondent group, isn't that evidence that the conditional distribution of V_j given U_j should also be similar to the conditional distribution of V_j given U_j in the respondent group? Or if the conditional distribution of some component of V_j given U_j is similar across all groups having both U_j and V_j recorded, isn't that evidence that the conditional distribution of that component of V_j given U_j in the jth group is similar to its distribution in the other groups?

In my experience, applied researchers often feel that the answers to these questions are "yes", and I think that their answers often reflect good judgement. Consequently, it appears to me that formalizing questions like these into hyperparameter models reflecting reasonable prior beliefs is an important research area demanding close cooperation between the applied researcher and the statistician. Once such hyperparameter models are formulated, the estimation task may demand substantial attention. However, I feel that having a collection of flexible imputation models reflecting a variety of reasonable prior assumptions is necessary for the sensible handling of nonresponse problems.

5. Acknowledgements

This work was largely supported by the Department of Health, Education, and Welfare as part of its planning effort for the proposed Survey of Income and Program Participation. It was facilitated by a Guggenheim Fellowship to the author. I would like to thank Tom Herzog, Dan Kasprzyk, and Fritz Scheuren for many helpful comments on this paper, as well as stimulating conversations about the real problems of nonresponse, especially as arising in the collection of income data in surveys.

REFERENCES

Basu, D. (1971), "An Essay on the Logical Foundations of Survey Sampling", in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston.

Buzen, J. P. (1975), "Operational Analysis: An Alternative to Stochastic Modeling", *Proceedings of the International Conference on the Performance of Computer Installations*, North Holland, June 1978.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977), *Foundations of Inference in Survey Sampling*, New York: John Wiley.

Cochran, W.G. (1977), *Sampling Techniques*, New York: John Wiley.

de Finetti, B. (1963), "Foresight: Its Logical Laws, Its Subjective Sources", in *Studies in Subjective Probability*, eds. H.E. Kyburg and H.E. Smokler, New York: John Wiley.

de Finetti, B. (1972), *Probability, Induction and Statistics*, New York: John Wiley.

Dempster, A.P. (1975), "A Subjectivist Look at Robustness", *Bull. I.S.I. Proc. 40th Session*, 349-374.

Dempster, A.P. (1977), "Examples Relevant to the Robustness of Applied Inferences", in *Statistical Decision Theory and Related Topics, II*, ed. S.S. Gupta and D.S. Moore, New York: Academic Press.

Diaconis, P. (1977), "Finite Forms of de Finetti's Theorem on Exchangeabilities", *Synthese 36*, 271-281.

Diaconis, P. and Freedman, D. (1978), "de Finetti's Generalizations of Exchangeability", *Stanford University Statistics Department Technical Report #109*.

Efron, B. and Morris, C. (1977), "Stein's Paradox in Statistics", *Scientific American, 236*, 5, 119-127.

Erickson, W.A. (1969), "Subjective Bayesian Models in Sampling Finite Populations"(with discussion), *Journal of the Royal Statistical Society, B, 31*, 195-233.

Feller, W. (1966), *An Introduction to Probability Theory and Its Applications, Volume II*, New York: John Wiley.

Fisher, R.A. (1935), *The Design of Experiments*, New York: Hafner.

Geisser, S. (1974), "A Predictive Approach to the Random Effect Model", *Biometrika, 61*, 1, 101-107.

Godambe, V.P. (1966), "A New Approach to Sampling from Finite Populations", *Journal of the Royal Statistical Society - B, 28*, 310-328.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory, Vols.I and II*, New York: John Wiley.

Hewitt, E. and Savage, L.J.(1955), "Symmetric Measures on Cartesian Products", *Trans. Amer. Mann. Society, 80*, 470-501.

James, W. and Stein, C. (1961),"Estimation with Quadratic Loss", in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, 361-379, Berkeley: University of California Press.

Kempthorne, O. (1952), *Design and Analysis of Experiments*, New York: John Wiley.

Lindley, D.V. (1972), *Bayesian Statistics, A Review*, Philadelphia: Society for Industrial and Applied Mathematics.

Madow, W.G. (1978), "Discussion of Papers by Basu, Royall and Cumberland", in *Survey Sampling and Measurement*, ed. N.K. Namboodiri, New York: Academic Press.

Royall, R.M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models", *Biometrika, 57*, 377-387.

Rubin, D.B. (1976), "Inference and Missing Data", Biometrika, 63, 3, 581-592 (with discussion and reply).

Rubin, D.B. (1977), "Formalizing Subjective Notions About the Effect of Non-Respondents in Sample Surveys", The Journal of the American Statistical Association, 72, 359, 538-543.

Rubin, D.B. (1977 unpublished manuscript), "The" Design of a General and Flexible System for Handling Non-Response in Sample Surveys", July 1.

Rubin, D.B. (1978 a), "Bayesian Inference for Causal Effects: The Role of Randomization," The Annals of Statistics, 6, 1, 34-58.

Rubin, D.B. (1978b), "The Phenomenological Bayesian Perspective in Sample Surveys from Finite Populations: Foundations", presented at the Spring Meetings of the Institute for Mathematical Statistics, Rutgers University, May 31, 1978.

Scott, A.J. (1977), "On the Problem of Randomization in Survey Sampling", Sankhya, Series 9, Volume 39, Pt. 1, 1-9.

Wald, A. (1950), Statistical Decision Functions, New York: John Wiley.