

# The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys

Donald B. RUBIN

## 1. INTRODUCTION

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 standard in practice, although often criticized by more mathematical statisticians who prefer to think about estimating parameters under some model.

I am very sympathetic with the imputation position. 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 we want to do since a hypothetical model is simply a structure that guides us to do sensible things with observed values.

Of course (1) imputing *one* value for missing datum can't be correct in general, and (2) in order to insert sensible values for a missing datum we must rely more or less on some model relating unobserved values to observed values. Hence, I see the best approach to be one where we can (1) insert more than one value for a missing datum, and (2) the inserted values reflect a variety of models for the dataset.

This position focusing on values to impute rather than parameters to be estimated is actually very Bayesian, and the Bayesian perspective guides us in our design of a general system for nonresponse. What we really want to impute is the "predictive distribution" of the missing values given then observed values (having integrated—averaged—over all model parameters). The theoretical Bayesian position tells us that (1) the missing data has a distribution given the observed data (the predictive distribution) and (2) this distribution depends on assumptions that have been made about the model. Notice that the (1)'s and (2)'s in the above paragraphs are meant to refer to the same two points.

The related practical questions are (1) how do we represent in a dataset a distribution of values to impute for each missing datum? And (2) what models should we use to tie observed and unobserved values to each other in order to produce the predictive distribution needed in (1). Section 2 addresses the first question and Section 3 addresses the second question.

## 2. REPRESENTING A DISTRIBUTION OF IMPUTED VALUES

Suppose that we have the predictive distribution for the missing values (i.e., algebraically derived under some model) or have an approximation to a predictive distribution (i.e., a large collection of units with complete data from which we can randomly choose for imputation). How to obtain such a distribution will be discussed later. The question addressed in this section is: how do we represent this distribution in the dataset? I see three ways.

### 2.1 Formally and Analytically Finding Predictive Distribution for Summary Statistics

The first way works only if we know exactly which summary statistics we would have calculated had there been complete data (and have some good statisticians to do Bayesian calculations). In this case, we really aren't interested in the individual missing data values but only in the values of the summary statistics based on observed and missing data. If we know (algebraically) the predictive distribution of the individual missing data points we can calculate (algebraically) the predictive distribution of the summary statistics without having to actually fill in any values.

My paper, "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys" (Rubin 1977), is the only example I know of where this is done. Although the calculations are explicitly worked through only for a very simple statistic (the sample average), a simple model for missingness, and a simple model for the distribution of the data (normal linear regression), the general idea is presented. Furthermore, some of the restrictions would be simple to eliminate: another linear function of the response variable (e.g., a regression coefficient or contrast between groups) is easy to handle, more general prior parameterizations are discussed and straightforward to incorporate, and polynomial regression terms can, of course, be incorporated.

The limitation of this approach is that calculating the predictive distribution of some summary statistics does not take care of the missingness problem in a general-purpose way (i.e., it does not necessarily help us with other data analyses we may wish to do). The advantage of the approach is that it is a correct analysis and so discussion can focus on the interpretation of the data and appropriateness of the model and not on the problems of the way values were imputed (i.e., the assumptions are clearly and precisely defined and there is no question about what the imputed values should be).

Ideally, even if we are to focus on one statistic such as the sample mean, we would try a variety of models to see how sensitive our inferences are to the models (e.g., quadratic terms in the regression model, different models of missingness).

### 2.2 Simulate Full Predictive Distribution via Many Imputed Datasets

A more generally applicable approach to representing a distribution of values for missing data is to generate many datasets drawn from the predictive distribution of the missing values, given the observed values. All these datasets agree for the observed data values; the values filled in for the missing values differ from dataset to dataset. Each dataset is filled in by taking

This work first appeared as Rubin, D. B. (1977), "The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys," prepared under contract for the U.S. Social Security Administration.

  *DOI: 10.1198/000313004X6355*

a random draw from the predictive distribution of missing data, given the observed data.

Generating many such datasets, analyzing each one, and summarizing how the conclusions change, is really the practical man's version of the method of Section 2.1: we simulate the predictive distribution of summary statistics instead of deriving it analytically. The method is more general than the methods of Section 2.1 in the sense that any analyses can be performed on the collection of filled-in datasets.

In some practical cases, we may be able to get a feeling for whether nonresponse creates a real problem by generating only a few such datasets, say five. If this is true, then when anyone wants to perform an analysis on "the" dataset, he should be required to perform it on all five filled-in datasets and at least eyeball how the results differ. A collection of five datasets is clearly not enough to estimate accurately the variability due to nonresponse, but it may be enough to see if the variability due to nonresponse is very large or very small compared to sampling standard errors calculated ignoring nonresponse.

This approach would be most appropriate with datasets having a limited number of units with nonresponse (since one would save, for each unit with missing data, five versions of his data) and analyses that are not expensive to perform.

### 2.3 Perform One Weighted Analysis With Several Versions of Missing Data

For some analyses, it may be possible to obtain information similar to that described in Section 2.2 without having to repeat the same analysis several times. The idea is to use the same type filled-in dataset as described above, but now we would include all five versions of each unit in one analysis: the weight for each unit with missing data would be split among the five versions of his filled-in data, and a weighted analysis would be done.

Beale and Little (1975) and Little (1976) propose such methods for multivariate normal data, and claimed to obtain good results under a simple model for missingness and using simple summary statistics. Such weighting and the corresponding analyses would presumably be most appropriate for datasets having a limited number of units with nonresponse, data analyses that can handle weighted observations, and datasets and analyses that are expensive to perform repeatedly (e.g., large total number of units). If the analyses are cheap, I would prefer to go with the method of Section 2.2 since it does in fact simulate a predictive distribution of summary statistics.

### 2.4 A Comment on System Requirements

In concluding this section, let me summarize some major points. First, if the data analysis to be performed is very common and will be used for many problems, it is best to develop an analytic tool to obtain the predictive distribution of the desired summary statistics of the analysis. Second, if actual imputations are to be made, it is wise to impute more than one value for each missing datum, and to perform the analysis several times using various filled-in values (or in some cases, perhaps one weighted analysis). Third, a general system must be flexible enough to handle multiple versions of each unit (preferably with and without weights) and usually to be able to perform the same analysis several times. Fourth, the data analyst must be willing to look at the results of the various data analyses and evaluate them remembering that (a) the filled-in values are based on some model and (b) some facets of the model are beyond confirmation by the data (the model for missingness). Note the extra burden on the user as well as the system.

## 3. OBTAINING THE PREDICTIVE DISTRIBUTION

Having outlined the sort of techniques that I would apply if I had a predictive distribution for missing data, it's now time to outline how to obtain such a distribution in practice. Any technique from imputing values will benefit from a large number of background variables that may be missing. The reason is rather obvious: having such variables implies that it's relatively easy to predict missing values, that is, the predictive distribution for a missing datum has small variance under reasonable models because knowing the values of background variables implies almost knowing the value of the missing datum.

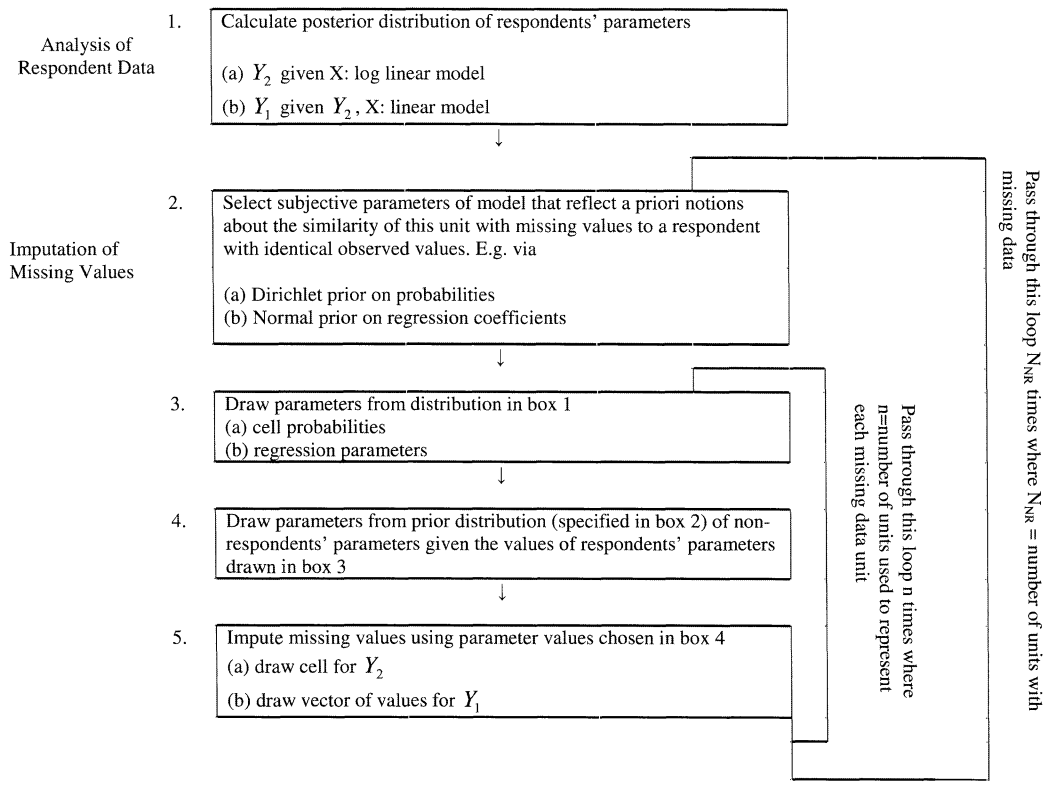
### 3.1 Restriction to Cases with Large Numbers of Respondents

At present, I think we should restrict attention to those surveys having a large number of respondents (units having essentially complete data) relative to the number of variables. This restriction allows us to ignore increased efficiency that may be available from using nonrespondents' data to help estimate respondents' parameters. If there do not exist a substantial number of respondents, we must face the missingness problem more directly: we must build one model for the distribution of the data for all units (not simply one model for the respondents and alternations on it for nonrespondents) and another model for the missingness process (the process that "causes" a unit to be a nonrespondent). Maximum likelihood model fitting, even assuming the missingness process is missing at random, can be difficult in general; computation with nonrandom missingness processes can be extremely involved. Future efforts should address this problem.

The current plan is to (1) model the respondent's data, (2) obtain for each nonrespondent, the conditional distribution of missing data (given observed data), as if he were a respondent (i.e., the predictive distribution under the model that says each nonrespondent is just like a respondent with the same values of observed variables), and (3) alter this distribution in various ways to allow for nonresponse bias that may change from nonrespondent to nonrespondent (e.g., the bias could change as type of nonresponse changes from refusal to not home or don't know). Step (3) reflects the missingness process and our fears that it may not be "missing at random" (Rubin 1976). By "obtain the distribution" in Step (2) I mean either an analytical expression or, more usually, a description of an algorithm for drawing observations from the distribution. Display 1 outlines this plan for imputing values. Each part of this outline is discussed in the following sections. Future work will explicate the computational details of each part of the plan.

### 3.2 Modeling the Distribution of the Data for Respondents

There are two commonly used models for multivariate data: (1) the normal and (2) the unrestricted multinomial. The normal is fine if it's a reasonable approximation because it uses a limited number of parameters; however, it may not be appropriate for all variables. The unrestricted multinomial is always appropriate, but it often fits too many parameters. Usually categories




*Display 1. Outline of plan for imputing values.*

must be collapsed, higher order interactions set to zero, etc. An important point to remember when fitting these models is that we do not necessarily have to model the full distribution of all variables. Assuming there are background variables recorded for all units, $X$, in order to produce the predictive distribution of missing values given observed values, we only have to model the conditional distribution of the other variables, $Y$, given $X$. (All symbols represent vectors.) For example, if the $Y$ variables are essentially continuous, we could easily fit the multivariate normal regression model for $Y$ given $X$. If some $Y$ are categorical and some $X$ are continuous, the conditional distribution of $Y$ given $X$ may be computationally difficult to fit. (Of course, logistic regression programs exist and can be incorporated into a system, but they are complicated compared to the usual linear regression programs.)

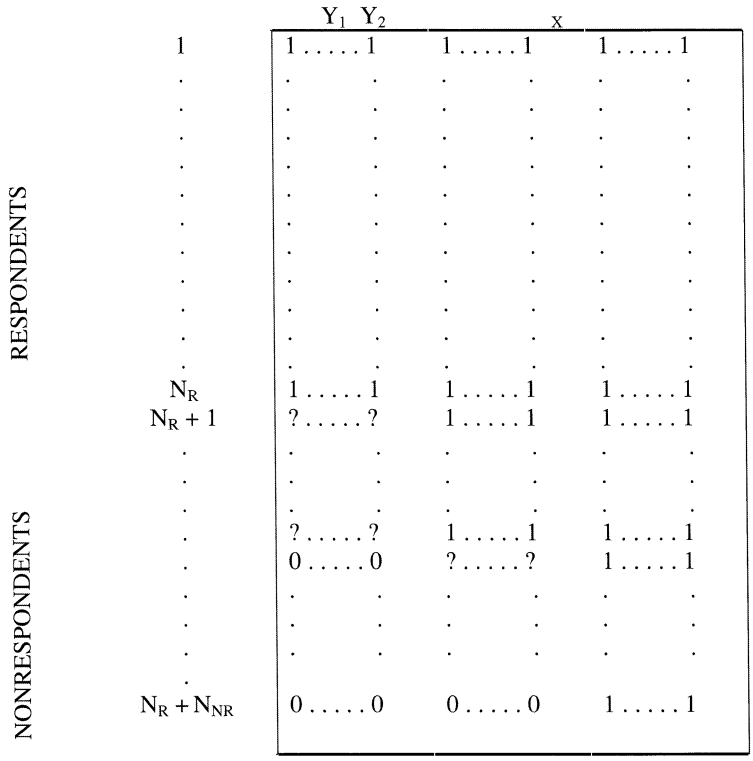
In order to rely only on generally available computer programs, the subsequent discussion will assume the following general case. Let $Y = (Y_1, Y_2)$ where $Y_1$ represents essentially continuous variables and $Y_2$ represents categorical variables; more specifically, $Y_1$ given $(Y_2, X)$ will be modeled as a multivariate normal linear regression and $Y_2$ given $X$ will be modeled as a conditional contingency table (i.e., the joint distribution of $(Y_2, X)$ can be represented by a contingency table). (Note: As we advance, we may want to include more robust methods of estimation, such as those that correspond to using median regression or reweighted least squares programs.) Now we assume that for all sampled units, $Y_2$ is "more observed" than $Y_1$, (Rubin 1974); that is, for each sampled unit, if any value in the vector $Y_1$ is observed, then all values in the vector $Y_2$ are observed. See Display 2 for the data structure. This restriction is not essential to the ideas being presented, but it does simplify discussion and reduce the need for exotic computer programs.

The problem of fitting a model for the complete-data units then reduces to (1) fitting a linear model to $Y_1$ given $Y_2$ and $X$ and (2) fitting a log-linear model to $Y_2$ given $X$. Methods for fitting linear models are much discussed; a good system should be aware of polynomial and interaction terms, as well as the issue raised by empirical Bayes arguments. Methods for fitting log-linear models were discussed, for example, by Bishop, Fienberg, and Holland (1975); again, empirical Bayes arguments are relevant.

Now, we don't really care about these models per se, but want to use them to impute values. Hence, we are not necessarily driven by the desire to estimate parameters well, but rather by the desire to predict unobserved values accurately. Empirical Bayes arguments are especially relevant in this case because we don't really care if our general model has "too many" parameters as long as they are smoothed in such a way that the smoothed predicted values are accurate. For some readable discussion of empirical Bayes methods, see the recent *Scientific American* article by Efron and Morris (1977).

### 3.3 Drawing Values to Impute Using the Models Developed for Respondents

Now assume that we have fit the models discussed in Section 3.2. Precisely, suppose that for the respondents we have the posterior distribution of the parameters of the log-linear $Y_2$ given $X$ model and the posterior distribution of the parameters of the normal linear regression $Y_1$ given $(Y_2, X)$ model; in most cases




*Display 2. Pattern of missing data.*

these posterior distributions would be independent because it is unlikely that we would want any a priori ties between these two sets of parameters.

Consider, in turn, each sampled unit with missing data. Data may be imputed independently for units under the usual model of IID $Y$ given $X$ and the parameters. Such models are also nearly completely general; see Diaconis (1976). Focus on any particular unit. Let his $Y = (Y_1, Y_2)$ be represented by $Y_1 = (Y_1^0, Y_1')$ and $Y_2 = (Y_2^0, Y_2')$ where the superscript 0 refers to missing values and the superscript $'$ refers to observed values. Then using the unit's observed values $X, Y_1', Y_2'$ and the models that have been fit for respondents, our job is to impute values for $Y_1^0, Y_2^0$ and we should be able to do this repeatedly.

Given the structure we have developed, this process is actually quite straightforward and proceeds sequentially. First assume the parameters of the models are known, that is, let the tow posterior distributions referred to above by point mass. First impute values for $Y_2^0$, say $\hat{Y}_2^0$, by drawing from the log-linear (or contingency table) model for $Y_2^0$ given $Y_2'$ and $X$; this is simply drawing a value from a multinomial with specified probabilities, that is, probabilistically picking the cell to which the observation belongs. Now, having $\hat{Y}_2^0$, imputed values for $Y_1^0$, say $\hat{Y}_1^0$, by drawing from the normal linear regression model of $Y_1^0$ on $Y_1'$, $Y_2'$, $Y_2^0$, $X$; this is simply drawing from a multivariate normal whose mean is specified by a known vector regression coefficient times the value of $(Y_1^0$ on $Y_1', Y_2', \hat{Y}_2^0, X)$ and whose covariance matrix is known and fixed for all values of $(Y_1', Y_2', Y_2^0, X)$. A simple way to obtain such a draw in practice is sequentially; draw the first component of $Y_1^0$ from the univariate normal given $(Y_1', Y_2', \hat{Y}_2^0, X)$, then draw the second component of $Y_1^0$ from the univariate normal given $(Y_1', Y_2', \hat{Y}_2^0, X)$ *and* the just-drawn value of the first component of $Y_1^0$, and so on.

Of course the drawings described in the preceding paragraph may be made repeatedly to generate several sets of imputed values for $(Y_1^0, Y_2^0)$.

Now let's consider how to reflect the fact that the posterior distributions for the parameters of these models are not a point mass. Before imputing values, draw the parameter values to be used in the imputation step from the posterior distribution of the parameters—a separate drawing from the posterior distribution for each set of imputed values. Note that imputing in the above described manner using stages of conditional distributions implies that the imputed values are in fact being drawn from the observed values assuming the missing data are missing at random.

Some approaches to imputing values do not explicitly build a model for respondents, for example, hot-deck, cold-deck procedures. These models find a group of "close" matches among the respondents for each unit with missing data; that is, for each unit with missing data, we redefine the category boundaries of a multinomial model for $X$ and the observed $Y$, $(Y_1', Y_2')$, to include his matches and himself. We then "estimate" the distribution for his missing $Y$, $(Y_1^0, Y_2^0)$, by drawing at random a unit among the matches to be used to fill in the missing data. Of course, if we are to fill in more than one set of values, more than one random drawing must be made. Generally, I feel that having a more explicit model (or several alternative explicit models) is better than such ad hoc methods because it allows us to borrow strength from complete-data units in a smooth way (the above method says either someone is a match or he is not and nonmatches are not used to estimate missing data—model building allows us to avoid these all-or-nothing decisions and to borrow strength more optimally). If there are a large number of matches for a narrowly defined band of matching characteristics, borrowing strength will be unimportant; otherwise, it is important. Furthermore, using explicit models means we know exactly the models and assumptions we are relying upon. Of course, it is likely that in many practical cases such sub-optimal procedures can serve as excellent expedients.

### 3.4 Altering the Predictive Distribution to Reflect Nonresponse Bias

The final part of our general approach is the ability to alter distributions to reflect nonresponse bias. Within the structure we have developed, this effort is straightforward, at least theoretically. In the last step of Section 3.3 we drew parameters from the posterior distribution of respondents' parameters; now we need to draw parameters from the posterior distribution of the nonrespondents' parameters rather than from the posterior distribution of the respondents' parameters.

This modified draw can be simply accomplished by adding a step between drawing the parameters and imputing values: given the value drawn for the respondents' parameters from the respondents' posterior distribution, now draw the nonrespondents' parameters from the prior distribution of the nonrespondents' parameters given the respondents' parameters with the chosen value of the respondent's parameters being the conditioned upon value. For the model of $Y_1$ given $Y_2$ and $X$ this prior could be simple a normal prior on the regression coefficients as in Rubin (1977); for the parameters of the model of $Y_2$ given $X$ this prior could be a Dirichlet prior (these are the conjugate priors).

The (subjective) parameters of these priors specify the data analyst's notions of nonresponse bias. Hence their values could (and usually should) change for different units in the sample depending upon reasons for nonresponse, subjective assessments recorded by interviewers but not included in $X$ or $Y$, and so on.

For each unit, each set of imputed value would arise from a new draw of both the respondents' and the nonrespondents' parameters. The five sets of imputed values would thus reflect the correct predictive distribution for missing data given observed data under the model for response bias.

### 3.5 A Note on the Required Effort

This document is only an overview of the work and an outline of an attack on the problem of nonresponse in sample surveys. However, each of the steps is well-defined and theoretically straightforward to carry out. Nevertheless, the implementation of many of the steps would require an enormous commitment of statistician time as well as systems and computer time. I believe that the effort would be exciting and the results very useful.

## REFERENCES


Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Series B, 37, 129–145.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Diaconis, P. (1976), "Asymptotic Expansions for the Mean and Variance of the Number of Prime Factors of a Number $n$," Tech. Report 96, Dept. of Statistics, Stanford University.

Efron, B., and Morris, C. (1977), "Stein's Paradox in Statistics," *Scientific American*, 237, 119–127.

Little, R. J. A. (1976), "Inference About Means for Incomplete Multivariate Data," *Biometrika*, 63, 593–604.

Rubin, D. B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

——— (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1977),"Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.