THE WEIGHTED SEQUENTIAL HOT DECK IMPUTATION PROCEDURE

Brenda G. Cox, Research Triangle Institute

One of the most commonly used item nonresponse adjustment procedures is hot deck imputation which uses current survey responses to substitute for missing data. In using the hot deck procedure, the sample individuals are usually partitioned into categories or imputation classes. Having formed these classes, the data records are then ordered according to important variables which influence response. Based upon the present data set or previous data, the hot deck procedure stores initial values by class for each variable to be imputed. As the survey data are processed, the imputation class to which an individual belongs is determined. If the record being processed is complete with respect to the variable or set of variables to be imputed, then that individual's responses replace the responses stored for the relevant imputation class of the hot deck. When a record is encountered with a missing response for an item, the last response stored in the hot deck for the same class is imputed for the missing response. When all records have been processed and the missing data imputed, estimates are usually computed without accounting for the effect of the imputation procedure.[1]

The procedure just described could be referred to as a sequential hot deck procedure since the data are first ordered and then the last reported value in the sequence is substituted for each missing value as the data are processed. The procedure is also unweighted in that the selection of a response for imputation purposes is independent of the sampling weight associated with the data record from which the response is taken and the data record to which a response is being imputed. However, ignoring sample weights implies that the distribution of responses within each imputation class of the imputation-revised data set may be distorted from that of the original distribution of responses.

Hot deck imputation is based upon the tacit assumption that nonrespondents would answer in a manner similar to that of respondents immediately adjacent to them in the sorted data file and hence that the data associated with the nearest neighbor is appropriate for imputation of missing values. For many surveys, this is not a reasonable assumption. In many instances, insufficient data are available to classify nonrespondents or the classifying data are only partially effective in explaining differences in response. When there are significant amounts of variation remaining within the imputation classes and when the sorting variables used within these classes are only partially effective in explaining this variation as well as differences in response rates, sample weights and probability selection of hot deck donors should be used in addition to the nearest neighbor approach to minimize imputation bias. In addition, many surveys oversample certain demographic groups so that one must utilize sample weights in implementing the imputation procedure to insure that the weighted distribution of responses for the imputation-revised data set reflects the weighted distribution of responses for the respondent data set.

1. Definition of the Weighted Sequential Hot Deck Imputation Procedure

By adapting a sequential sample selection method discussed by Chromy (1979), a weighted sequential hot deck algorithm has been developed for use in imputing missing data.[2] This algorithm is designed so that means and proportions, estimated using the imputation-based data set, will be equal in expectation to the weighted mean or proportion estimated using respondent data only. Noting that variances, covariances, correlations, regression coefficients, and other higher order population parameters are estimated by simple functions involving weighted means of products and cross-products, the expectation of such higher order statistics over repeated imputations should also equal the corresponding estimator obtained from respondent data only. Achieving this desirable result implies that the weights associated with respondents and nonrespondents must be considered so as not to seriously bias the estimated distribution of variates obtained from respondent data.

An objection to the usual unweighted procedure is that the data from a respondent can be used many times for imputation since the data from a respondent record will be imputed to every nonrespondent record following in the data file until another respondent record is encountered. Further, because of their position in the sorted data file, many respondent records will be arbitrarily excluded from "donating" data to nonrespondent records. The weighted sequential hot deck procedure has the additional advantage that it controls the number of times that a respondent record can be used for imputation and gives each respondent record a chance to be selected for use as a hot deck donor.

The imputation strategy may be thought of as utilizing two data files, a data file of respondents and a data file of nonrespondents. The imputation procedure involves using data for responding individuals to substitute for the missing data associated with nonresponding individuals. The first step in the proposed strategy is to sort the two data files with respect to variables related to response and the distribution of characteristics of interest. Both files will have sampling weights attached to each individual. The imputation will occur within imputation classes so that the distribution of means and proportions for discovered visits will be preserved within each class over repeated imputations.

In sequential order, the data associated with respondents are assigned to nonrespondents. The number of times that the data record for respondent i is accessed to impute to nonresponding individuals will be denoted by n(i). The sum of the n(i) over all responding individuals must equal the number of nonresponding individuals to which data are to be imputed. This

will be true since the data record for one, and only one, respondent is accessed for imputation to each nonrespondent. Thus, if r equals the total number of responding individuals and if n equals the total number of nonrespondents, then

$$\sum_{i=1}^{r} n(i) = n \quad .$$

The number of times the data for respondent i are imputed for the missing data of nonrespondents will be a function of the sampling weight of respondent i and the sampling weights of the nonrespondents to which it can potentially be imputed.

Some additional terms must be defined in order to describe the imputation algorithm. Assume the units in the responding and nonresponding data files are numbered from 1 to r and 1 to n respectively. Then let s(i) be the sampling weight attached to the i-th element of the respondent data file (i=1,2,$\cdots$,r) and w(j) be the sampling weight attached to the j-th element of the nonrespondent data file (j=1,2,$\cdots$,n). The sampling weights used with respondents and nonrespondents need not be defined differently. However, when subsampling of nonrespondents occurs, the sampling weight s(i) attached to respondents should be based upon weights adjusted for the subsampling of nonresponding individuals.

For each element of the nonrespondent data file, define the term v(j) to be

$$v(j) = w(j)\, s(+)/w(+)$$

where s(+) is the sum of the respondent weights and w(+) is the sum of the nonrespondent weights; that is,

$$s(+) = \sum_{i=1}^{r} s(i)$$

$$w(+) = \sum_{j=1}^{n} w(j) \quad .$$

The imputation algorithm can be thought of as partitioning the respondent data file into n zones of variable width v(j) where j=1,2,$\cdots$,n and then imputing the response for the j-th nonrespondent from a respondent in the corresponding zone of the respondent data file. The width of the zone v(j) reflects the effect of the weight of nonrespondent j in relation to the sum of the weights attached to all nonrespondents.

The imputation algorithm actually proceeds in the following fashion. Define I(i) to be the integer such that

$$\sum_{j=1}^{I(i)+1} v(j) > \sum_{i'=1}^{i} s(i') \geq \sum_{j=1}^{I(i)} v(j)$$

and F(i) to be

$$F(i) = \left[\sum_{i'=1}^{i} s(i') - \sum_{j=1}^{I(i)} v(j)\right]/v[I(i) + 1]$$

The imputation algorithm will be such that the minimum number of times the i-th response can be used for imputation is I(i) - I(i-1) and the maximum is I(i) - I(i-1) + 1. By definition, I(0) = 0 and F(0) = 0. The quantity I(i) arises from the restriction that the weights attached to the first i respondent records (i=1,2,...,r) will determine the minimum number of times these i records will be used for imputation. Thus the algorithm stipulates that the data from the first i respondent records must be used to impute data to the first I(i) nonrespondents. The probability that data from the first i respondents will be used to impute data to the first I(i) + 1 nonrespondents is equal to F(i). Essentially one can say that the weights attached to the first i respondent records demand that these records be used I(i) + F(i) times for imputation where I(i) is the integer number of times and F(i) is the fractional remainder.

With this in mind, the imputation algorithm can be described as a sequential selection scheme [n(i) or the number of times the i-th response is to be used for imputation is being selected] where

$$P(i) = \text{Prob}\left\{\sum_{i'=1}^{i} n(i') = I(i) + 1 \,\middle|\, \sum_{i'=1}^{i-1} n(i')\right\}$$

is defined as a function of F(i) and F(i-1). The variable P(i) is the conditional probability that the first i respondent records will be used I(i) + 1 times for imputation given the number of times that the first i-1 records were used for imputation. Conditional sequential probabilities of selection for the three possible cases are presented in Table 1.

To select the number of times each respondent record is to be used for imputation,

$$\sum_{i'=1}^{i} n(i')$$

is determined for each value of i by this calculation of the appropriate conditional probability and comparison of the probability with a uniform random number. If the random number is less than the appropriate conditional probability,

$$\sum_{i'=1}^{i} n(i') \quad ,$$

is set to I(i) + 1; otherwise, it is set to I(i). The values of n(i) are determined as

$$n(i) = \sum_{i'=1}^{i} n(i') - \sum_{i'=1}^{i-1} n(i')$$

Having determined the number of times each respondent's record will be used for imputation, missing data will be imputed as follows. The first n(1) nonrespondents will have their missing data replaced by data imputed from the first respondent. The next n(2) nonrespondents will have data imputed from the second respondent and so on until the last n(r) nonrespondents have their missing data replaced by data imputed from the r-th respondent (the last respondent) in the

respondent data file. The procedure described in this section is illustrated in Figure 1.

2. Properties of the Weighted Sequential Hot Deck Imputation Procedure

The weighted sequential hot deck procedure was constructed to insure that over repeated imputations the data imputed to nonrespondents would have the same mean value as that which can be estimated from the respondent data using appropriate weights. This may be symbolized in the following manner. Define the mean of the imputed data to be $\bar{Y}$ where

$$\bar{Y} = \sum_{j=1}^{n} w(j)\, Y(j)/w(+)$$

and Y(j) is the response imputed to the j-th nonrespondent. Further, define the estimate of the mean which can be obtained from the respondent data to be $\bar{X}$ where

$$\bar{X} = \sum_{i=1}^{r} s(i)\, X(i)/s(+)$$

and X(i) is the response given by the i-th respondent. Using this notation, the weighted sequential hot deck procedure was designed so that

$$E_I\,[\bar{Y}] = \bar{X} \qquad (1)$$

where the expectation is taken over all possible imputations.

To prove that this property holds, note that one may express the average value imputed to the nonrespondents as

$$E_I\,(\bar{Y}) = E_I\,[\sum_{i=1}^{r} w'(i)\, X(i)/w(+)] \qquad (2)$$

where

$$w'(i) = \sum_{j=1}^{n} \lambda_i(j)\, w(j) \qquad (3)$$

and

$$\lambda_i(j) = \begin{cases} 1 & \text{if the j-th nonrespondent has data imputed from the i-th respondent,} \\ 0 & \text{otherwise.} \end{cases}$$

Since one and only one response is to be imputed to each nonrespondent,

$$w'(+) = w(+) \quad .$$

Note that one can show that equation (1) is true when the expression w'(i) in equation (2) has expectation over all imputations of

$$E_I\,[w'(i)] = s(i)\, w(+)/s(+) \quad . \qquad (4)$$

To prove equation (4), we will first show that

$$E_I\,[\sum_{i'=1}^{i} w'(i')] = \sum_{i'=1}^{i} s(i')\, w(+)/s(+) \qquad (5)$$

is true for all i=1,2,$\cdots$,r and then by simple subtraction it is clear that equation (4) is true.

In order to prove these statements, the following lemma is needed.

Lemma: After each sequential imputation step, the following two conditions hold:

$$\text{Prob}\,\{\sum_{i'=1}^{i} n(i') = I(i) + 1\} = F(i)$$

$$\text{Prob}\,\{\sum_{i'=1}^{i} n(i') = I(i)\} = 1 - F(i) \quad .$$

This lemma can be proved inductively since the algorithm is applied sequentially. All three deterministic conditions must be considered at each step of the proof.

In order to prove that the weighted sequential hot deck procedure imputes data in such a manner that the mean of the imputed data is equal in expectation to the mean estimated from respondent data only, the first step will be to show that equation (4) is true for i=1 or that

$$E_I\,[w'(1)] = s(1)\, w(+)/s(+) \quad .$$

Note that according to the above lemma

$$\text{Prob}\,\{n(1) = I(1)\} = 1 - F(1)$$

and

$$\text{Prob}\,\{n(1) = I(1) + 1\} = F(1)$$

and thus since imputation occurs sequentially we have

$$E_I\,[\lambda_1(j)] = \begin{cases} 1 & \text{for } j=1,2,\cdots,I(1) \\ F(1) & \text{for } j=I(1)+1 \\ 0 & \text{otherwise.} \end{cases}$$

The expectation of w'(1) over all imputations is

$$E_I\,[w'(1)] = \sum_{j=1}^{I(1)} w(j) + F(1)\, w[I(1)+1] \quad .$$

Recalling that the v(j) are defined to be

$$v(j) = w(j)\, s(+)/w(+) \quad ,$$

we have

$$E_I\,[w'(1)] = \sum_{j=1}^{I(1)} v(j)\,[w(+)/s(+)] + \{F(1)\, v[I(1)+1]\, w(+)/s(+)\} \quad .$$

By definition

$$F(1) = [s(1) - \sum_{j=1}^{I(1)} v(j)] \,/\, v[I(1)+1] \quad ,$$

so that

$$E_I\,[w'(1)] = s(1)\, w(+)/s(+) \quad .$$

We have now shown that equation (4) is true for i = 1.

The remainder of the proof will be by induction.

Suppose for i'=1,2,...,i-1,

$$E_I\,[w'(i')] = s(i')\, w(+)/s(+)$$

and hence

$$E_I \left[\sum_{i'=1}^{i-1} w'(i') \right] = \sum_{i'=1}^{i-1} s(i')\, w(+)/s(+) \quad . \quad (6)$$

Using equation (3), we may write

$$\sum_{i'=1}^{i} w'(i') = \sum_{i'=1}^{i} \sum_{j=1}^{n} \lambda_{i'}(j)\, w(j) \quad .$$

One can again use the lemma to show that since the imputation procedure proceeds sequentially,

$$E_I \left[\sum_{i'=1}^{i} \lambda_{i'}(j) \right] = \begin{cases} 1 & \text{for } j=1,2,\ldots,I(i) \\ F(i) & \text{for } j = I(i+1) \\ 0 & \text{otherwise} \end{cases}$$

and hence

$$E_I \left[\sum_{i'=1}^{i} w'(i') \right] = \sum_{i'=1}^{I(i)} w(i') + F(i)\, w[I(i)+1]$$

Using the definitions

$$w(j) = v(j)\, w(+)/s(+)$$

and

$$F(i) = \left[\sum_{i'=1}^{i} s(i') - \sum_{j=1}^{I(i)} v(j) \right] / v[I(i)+1] \ ,$$

we have

$$E_I \left[\sum_{i'=1}^{i} w'(i') \right] = \sum_{i'=1}^{i} s(i)\, w(+)/s(+) \quad . \quad (7)$$

Using (6) and subtracting $\sum_{i'=1}^{i-1} w'(i')$ from equation (7), we have

$$E_I [w'(i)] = s(i)\, w(+)/s(+)$$

which by induction is true for $i=1,2,\cdots,r$. This completes the proof verifying that the weighted sequential hot deck procedure, over repeated imputations, imputes data to nonrespondents in such a manner that the mean of the imputed data is equal in expectation to the mean of the respondent data.

3. Applications of the Weighted Sequential Hot Deck Imputation Procedure

The weighted sequential hot deck imputation procedure was developed for use in a double sampling context for the National Medical Care Expenditure Survey (NMCES). The purpose of this survey is to provide detailed information on the health of the residents of the United States, how and where they receive health care, the cost of the services, and how these costs are paid. The report from this study will ultimately have an impact on public policy concerning health care for the entire nation.[3]

Much of this data could only be obtained in a household interview. For this reason, NMCES selected a national sample of 13,500 households for interview during the 1977 calendar year. Since estimates of health care utilization and expenditures obtained from household-reported data are known to be subject to bias, NMCES also included a record check component in which a one-third subsample of the NMCES sample individuals had their medical providers surveyed as a part of the Medical Provider Survey (MPS).[4] Hence, using NMCES household data together with MPS provider data, one can develop double sampling estimates for health care variables.

For ease in analysis of the resultant household and provider data, it was decided that expenditure and utilization data, as the provider would have reported it, were to be imputed to all NMCES sample individuals with this data missing. The specification was made that estimates obtained using the provider-reported data, whether imputed or real, were to be equal in expectation over repeated imputations to the conventional double sampling estimate. It was in this context that the weighted approach for hot deck imputation was developed. Note that, for this particular example, the usual assumption made in imputing data--that the distribution of responses for the nonrespondents is similar to that of the respondents--is true since the "respondents" are a probability sample selected from the full NMCES sample.

Presently, the software for Chromy's sequential sample selection algorithm has been revised so that it permits the weighted sequential selection of hot deck donor records.[6] Basically, this involves the calculation of a pseudoweight which is then used in conjunction with the Chromy-Williams software for sequential sample selection.[5] These pseudoweights are defined sequentially as follows:

$$S^*(1) = I(1) + F(1)$$
$$S^*(2) = I(2) + F(2) - S^*(1)$$
$$S^*(3) = I(3) + F(3) - S^*(2)$$
$$\vdots$$
$$S^*(r) = n - S^*(r-1).$$

Note that by definition $F(r) = 0$ and $I(r) = n$ where r is the total number of respondents. When these pseudoweights are used with the Chromy-Williams sequential sample selection software, the software yields $n(i)$ for each of the $i = 1,2,\ldots,r$ respondents as specified by the weighted sequential hot deck procedure. Records with missing data (recipient records) must then be processed sequentially with the first $n(1)$ recipient records receiving their data from the record corresponding to the first respondent (the first donor record). The next $n(2)$ recipient records receive data from donor record number 2 and so on until the last $n(r)$ recipient records have data imputed from the last or r-th donor record.

The disadvantage of the weighted sequential hot deck technique is that it requires four passes through the data in order to impute values; the first pass sorts the file, the second pass defines pseudoweights, the third pass determines the $n(i)$, and the fourth pass imputes missing data. The unweighted procedure requires only two passes as it sorts the file and then imputes data directly.

As a part of the MPS project, the effect of weighted sequential hot deck imputation on survey estimates is being evaluated. To reflect the double sampling nature of the imputation, a test data set has been created composed of all individuals participating in the provider check survey. This test data set will be subsampled so as to partition the data set into two parts. From the individuals in one part of the test data set, the provider check data will be ignored. Data will be independently imputed to them two times. The entire subsampling and imputation operation will be repeated three times so that the double sampling situation is replicated three times and the imputation two times for each of the three partitions of the test data set. The estimates obtained using the imputation-revised data sets will be compared with the estimates obtained using the actual data. This approach is being used since it will allow the estimation of the variance due to the subsampling as well as the imputation variance. The results of the evaluation of the effect of imputation on survey estimates involving health care expenditures and utilization variables will be reported at a later date.

Presently, tentative plans exist for using the weighted sequential imputation procedure in a cold deck approach to impute provider-reported data. The National Medical Care Utilization and Expenditure Survey (NMCUES), a survey similar to NMCES, is being conducted this year. In lieu of a provider check survey for NMCUES for which funds are not available, it is planned that provider check data from NMCES and the relationships that exist between household and provider reported data for that survey will be used to impute provider-reported data for visits reported in NMCUES. Depending on the results of the evaluation studies conducted for NMCES, provider-reported expenditure and utilization data may be imputed for visits reported in the NMCUES survey.

4. Concluding Remarks

The weighted sequential hot deck imputation approach has important properties which will make it preferrable in many instances. When the response rate is low, a weighted approach is intuitively more acceptable than the conventional approach which has much less control over how often a respondent's data are used for imputation or over how the distribution of the imputation-derived data differs from the distribution of respondent data. Further, when the sorting and classifying variables explain only a small portion of the variation in response (which is frequently the case), it becomes much more important that imputation not disturb the observed distribution of responses. The weighted approach is more costly than traditional methods but is well worth the cost when the response rate is not high or when the classifying variables are only partially effective in explaining response.

5. Acknowledgements

The author would like to acknowledge the advice and assistance received from Ralph E. Folsom in preparing this paper. It was Mr. Folsom who suggested that Chromy's sequential sample selection algorithm could be adapted to create a weighted hot deck imputation algorithm.

The research in this paper was done for the National Center for Health Services Research (NCHSR) under Contract No. HRA-230-76-0268. The views expressed in this paper are those of the author and no official endorsement by NCHSR is intended or should be inferred.

6. Footnotes

1 A general description of the hot deck imputation procedure and other nonresponse imputation procedures is given in Chapman, 1976.

2 Sequential sample selection methods are distinguished by the manner in which random numbers are used to determine the sample. These methods require that each sampling unit in the frame be considered in order and a probabilistic decision made concerning its inclusion in the sample. Chromy's 1976 paper extends the concept of sequential sample selection to more general unequal probability sampling schemes.

3 The sample design and weighting procedures used for NMCES are discussed in Cox 1980b.

4 The sample design and weighting procedures used for MPS are discussed in Cox 1980a.

5 The software for sequential sample selection is discussed in Williams and Chromy (1980).

6 Tracy Duggan and Vincent Iannacchione at the Research Triangle Institute have revised the Chromy-Williams software so that it does multiple weighted sequential hot deck imputation.

7. References

[1] Chapman, David W. (1976). A survey of nonresponse imputation procedures. American Statistical Association 1976 Proceedings of the Social Statistics Section, 324-329.

[2] Cox, Brenda G. (1980a). Construction of Sample Weights for the Medical Provider Survey. RTI Report No. RTI/1320/44-03W. Prepared for the Health Resources Administration under Contract No. HRA-230-76-0268.

[3] Cox, Brenda G. (1980b). Development of Sample Weights for the National Medical Care Expenditure Survey. RTI Report No. RTI/ 1320/04-01F. Prepared for the Health Resources Administration under Contract No. HRA-230-76-0268.

[4] Chromy, James R. (1979). Sequential sample selection methods. American Statistical Association 1979 Proceedings of the Survey Methodology Section.

[5] Williams, Rick L. and James R. Chromy (1980). SAS sample selection macros. Proceedings of the Fifth Annual SAS Users Group International Conference.

Table 1. Determination of the Conditional Sequential Probabilitis of Selection

Deterministic Condition	Value of P(i)	
	$\sum_{i'=1}^{i-1} n(i') = I(i-1)$	$\sum_{i'=1}^{i-1} n(i') = I(i-1) + 1$
$F(i) = 0$	0	0
$F(i) > F(i-1) \geq 0$	$[F(i)-F(i-1)]/[1-F(i-1)]$	1
$0 < F(i) \leq F(i-1)$	0	$F(i)/F(i-1)$

Figure 1. Example of Data Files for a Weighted Sequential Hot Deck Procedure

Respondent Data File

Record No. (i)	Adjusted Sample Weight s(i)	$\sum_{i'=1}^{i} s(i)$	I(i)	F(i)
1	10	10	0	0.29
2	20	30	0	0.86
3	5	35	1	0.00
4	15	50	1	0.30
5	12	62	1	0.54
6	13	75	1	0.80
7	14	89	2	0.05
8	9	98	2	0.17
9	7	105	2	0.27
10	15	120	2	0.47
11	18	138	2	0.71
12	16	154	2	0.92
13	13	167	3	0.18
14	16	183	3	0.58
15	9	192	3	0.80
16	8	200	4	0.00

Nonrespondent Data File

Record No. (j)	Sample Weight w(j)	v(j)	$\sum_{j'=1}^{j} v(j)$	Data Can Be Imputed From Respondents
1	7	35	35	1,2,3
2	10	50	85	4,5,6,7
3	15	75	160	7,8,9,10,11,12,13
4	8	40	200	13,14,15,16