COMPARISON OF TWO PROCEDURES FOR IMPUTING MISSING SURVEY VALUES

John C. Bailar, III, National Cancer Institute, Barbara A. Bailar, U.S. Bureau of the Census

Summary

Survey data may be viewed as a matrix in which columns correspond to respondents and rows to specific survey items. Missing values may be imputed from functions of other items in the same column (a class that includes most regression methods), or in the same row (a class that simply ignores missing values as well as hot deck procedures), or in both columns and rows. Hot deck procedures have been widely used for over 25 years but their properties are almost unexplored. This paper compares the first two moments of estimated row means when missing values are a) ignored, or b) imputed by hot deck procedures. Under usual and/or reasonable assumptions both methods are unbiased. Their relative variances depend on the correlation structure of the data within rows and on the positions of the missing values. Exact formulas and limiting cases will be presented for $\rho_{ij} = \rho^{|i-j|}$ and/or $\rho_{ij} = 0$ unless $|i-j| \leq 1$.

(Note: Tables, figures, and appendices are available from the authors.)

Introduction

Consider a random sample (simple random or otherwise) of size n, drawn from an infinite r-variate population. The data may be considered to form an rxn matrix of observations. Assume that some variables (say 1, 2, ...,q) are always observed and are used to stratify the population in any suitable way. We deal only with sample items within strata, so these items play no further role in the analysis here. For the other variates, q+1, q+2,...,r, some observations may be missing, while others are identified as outliers or are otherwise not available for analysis. The problem is to estimate the population mean, μ, for any one of these variates or, equivalently, to estimate $n\mu = T$, where n is the sample size.

For specificity, assume that variate t is missing for m of the n sample elements. Several different approaches have been developed to deal with the problems posed by the missing observations. These include:

1. Equal-weights models, in which the weight initially assigned to each missing variate is redistributed equally over the sample elements in which that variate was actually observed. This corresponds in important respects to a procedure of ignoring missing values and defining the sample size for a specific variate as equal to the number of times that variate was observed.

2. Regression models, in which missing values for any sample element are estimated by regression methods from the variates that were observed for that element; the estimated value is then used as if it had been observed, with appropriate adjustments to the variance. In rough terms, this means that one estimates missing values in any column of the rxn data matrix from other values in that column.

3. Hot-deck models, in which missing variates for any sample element are estimated by a linear combination of the values for other sample elements for which that variate was observed; this is a procedure for estimating missing values in any row of the rxn data matrix from other values in that row. This approach in general is not suitable for simple random samples, since it is critically dependent on the ordering of the sample vectors (the columns of the data matrix). In the hot-deck model considered in this paper, a missing variate is considered to equal the immediately preceding value actually observed for that variate; to accomodate missing values in the first sample vector a complete initial (cold-deck) vector is established by some means before the data are processed.

4. An extension of the models in which missing values are estimated by using the complete data matrix (both rows and columns). There seems to have been no specific application of such methods to survey data though much has been done in the field of experimental design.

Each of the first three approaches has its own advantages and disadvantages, of which we give only a few examples. Equal-weights models are conceptually and computationally simple but generally require the assumption that missing values are a random subset of all values. They may lead to serious loss of data in cross-tabulations and they may be much less efficient than other models. Regression models seem especially appropriate when variates likely to be missing are closely correlated with other items more likely to be present in the same sample vector. However, they can become complicated and unwieldly if the set of missing variates differs widely from one vector to another. Hot-deck models seem to have some advantages when the successive sample elements are correlated. This might be as a result of the sequence in which data are submitted (a block of data from rural residents, then a block from city residents, etc.), or a serial correlation could be deliberately induced in the data by appropriate sorting of the sample vectors prior to processing.

The equal-weights model will be compared with a version of the hot-deck model used in the Current Population Survey (CPS), a large probability survey conducted monthly by the Bureau of the Census.

It is convenient to assume that the data matrix is complete, but is paired with another rxn matrix of elements that are zero or one, depending on whether the corresponding item in the data matrix was missing or observed. This matrix is called the window matrix. From here we consider only one vector (row) at a time from the data matrix, designated $\underline{x}$, and the corresponding vector from the window matrix, designated $\underline{w}$. Assume that a cold-deck has been established, so that $\underline{x}$ and $\underline{w}$ each have n+1 elements, labeled 0, 1, 2,..., n.

It is now assumed that $\underline{x}$ and $\underline{w}$ are independent; this implies that whether an item is missing is independent of its value. This is a strong assumption with respect to many situations in which missing values must be estimated. However, we do not yet assume independence of the elements within $\underline{x}$ and $\underline{w}$. The value of the population mean, μ, is now estimated by a linear combination of the elements of $\underline{x}$, $n\hat{\mu} = \sum_{i=0}^{n} c_i x_i$ where each c_i is a function of $\underline{w}$ with $c_i = 0$ when $w_i = 0$ and $\sum_{i=0}^{n} c_i = n$. The values $c_0, c_1, \ldots, c_n$ now form a third random (n+1) vector $\underline{c(w)}$ for each row of the data matrix.

The problem may be seen as that of comparing two algorithms for constructing $\underline{c}$ from $\underline{w}$. In the case of the equal-weights procedure, $w_0 = c_0 = 0$ and $c_i = n/(n-m)$ for each value observed. In the case of the hot-deck procedure used in the CPS, c_i for each value observed in the sample (i=1, 2,...,n) is 1 plus the length (perhaps zero) of the string of missing values immediately following element i, while c_0 is 1 less than this. For both procedures, it is assumed that all x_i, i=0,...,n have the same marginal distribution with mean μ and variance σ^2.

It is easily shown that $\hat{\mu}$ is unbiased under any set of weights $\underline{c}$ compatible with the assumptions used here. Thus it is reasonable to compare the two procedures in terms of $\mathrm{Var}(\hat{\mu})$. These will be designated $\mathrm{Var}_{EW}(\hat{\mu})$ and $\mathrm{Var}_{HD}(\hat{\mu})$ for variances under the equal-weights and hot-deck procedures, respectively.

It can also be easily verified, under the assumptions stated, that with any set of weights $\underline{c}$

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n^2} \sum_{i=0}^{n} \sum_{j=0}^{n} \mathrm{Cov}(c_i x_i, c_j x_j)$$

$$= \frac{1}{n^2} \sum_{i=0}^{n} \sum_{j=0}^{n} E(c_i c_j)\mathrm{Cov}(x_i x_j) \quad .$$

(See Appendix 1)

In some applications $\mathrm{Var}(\hat{\mu})$ will be found conditional on the observed vector $\underline{w}$; i.e. conditional on $\underline{c}$, in which case $E(c_i c_j) = c_i c_j$. However, unless otherwise noted, the following development is in terms of expectations over $\underline{c}$. It is assumed that n, the sample size, and m, the number of observations missing from the sample of initial size n, are fixed and known. We now assume also that all possible arrangements of the m missing values are equally likely, so that $p(w_i = 1) = (n-m)/n$ and for $i \neq j$

$$P(w_i = w_j = 1) = \frac{n-m}{n} \cdot \frac{n-m-1}{n-1} \quad .$$

Equal-Weights Procedure

In equal-weights procedures $\underline{c}$ is constructed by the following algorithm: $c_i = 0$ if $w_i = 0$ and $c_i = n/(n-m)$ if $w_i = 1$ for i=1, 2, ..., n. We consider three possible functions for $\mathrm{Cov}(x_i, x_j)$. The first covariance structure is that for simple random sampling, in which the elements of $\underline{x}$ are independent and identically distributed. Then

$$\mathrm{Var}_{EW}(\hat{\mu}) = \frac{1}{n^2} \sum_{i=0}^{n} E(c_i^2)\sigma^2 = \sigma^2/(n-m) \quad .$$

For the second example, assume that the observed values of $\underline{x}$ are correlated, with $\mathrm{Cov}(x_i, x_j) = \sigma^2 \rho^{|i-j|}$ for $i \neq j$. We now need, for $i \neq j$, $E(c_i c_j)$. Since $c_i c_j$ is zero unless c_i and c_j both equal $n/(n-m)$,

$$E(c_i c_j) = \left(\frac{n}{n-m}\right)^2 \cdot P_r\left[c_i = c_j = \left(\frac{n}{n-m}\right)\right]$$

$$= \left(\frac{n}{n-m}\right)^2 \frac{n-m}{n} \cdot \frac{n-m-1}{n-1}$$

$$= \frac{n(n-m-1)}{(n-m)(n-1)} \quad .$$

Then

$$\sum_{j=2}^{n} \sum_{i=1}^{j-1} E(c_i c_j)\mathrm{Cov}(x_i x_j)$$

$$= \frac{n(n-m-1)}{(n-m)(n-1)}\sigma^2 \sum_{j=2}^{n} \left(\frac{\rho - \rho^j}{1-\rho}\right)$$

$$= \frac{n(n-m-1)}{(n-m)(n-1)}\, \sigma^2 \left(\frac{\rho}{1-\rho}\right)\left(n - \frac{1-\rho^n}{1-\rho}\right)$$

so that

$$\mathrm{Var}_{EW}(\hat{\mu}) = \frac{\sigma^2}{n-m} + \frac{(n-m-1)}{(n-m)(n-1)}\left[\frac{2\sigma^2\rho}{n(1-\rho)}\right]\left(n - \frac{1-\rho^n}{1-\rho}\right) .$$

One can show that $\mathrm{Var}_{EW}(\hat{\mu})$ is strictly increasing in ρ over the range $\rho \in [-1, 1]$.

(See Appendix 2)

For the third covariance structure, assume that $\mathrm{Cov}(x_i, x_j) = \rho\sigma^2$ for $i \neq j$. In this case

$$\mathrm{Var}_{EW}(\hat{\mu}) = \frac{\sigma^2}{n-m} + \frac{\rho\sigma^2}{n} \sum_{i \neq j}^{n} \frac{(n-m-1)}{(n-m)(n-1)}$$

$$= \frac{\sigma^2}{n-m}\left[1 + (n-m-1)\rho\right] \quad .$$

Hot-deck Procedure: Properties of $\underline{c}$

The procedure used to impute missing values in the CPS is computationally simple, but its properties are almost unexplored (Bailar, Bailey, and Corby, Survey Sampling and Measurements, Chapter 12, November 1978). In this procedure, a "cold-deck" value x_0 is established before the data are examined, where x_0 is assumed to be an observation from the same population as the other elements of $\underline{x}$ but independent of them. Commonly, x_0 is taken from the same stratum of a previous survey. The new data are then processed in the sequence in which they are submitted, which is generally non-random with respect to the values in $\underline{x}$. If the element x_i is observed, it is used both in the computations and to replace the value in the cold-deck (which is now called a hot-deck). If x_i is not observed, the value currently in the cold-deck (initially) or the hot-deck (after the first observed value replaces x_0) is used in the computations.

The effect of this procedure is to use each observed value x_i a random number of times c_i. This is clearly not an optimal procedure in the case of simple random sampling. However, when successive elements of $\underline{x}$ are positively correlated one might expect each missing value to be replaced (imputed) by a value closer to it than the sample mean.

It may be helpful to illustrate the computation of $\underline{c}$ in this procedure. Assume that n=12 (so that $\underline{x}$, $\underline{w}$, and $\underline{c}$ each have 13 elements) and that observations 1, 2, 5, 6, and 9 are missing. Then

$\underline{w}$ = (1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1)

$\underline{c}$ = (2, 0, 0, 1, 3, 0, 0, 1, 2, 0, 1, 1, 1)

Since we have assumed that m, the number of missing observations, is fixed, and that each possible arrangement of them is equally likely except that x_0 is always present, we can find the marginal and joint probability distributions of the elements of $\underline{c}$. Note that $P(c_0 = 0, c_1 = 0) = 0$ since $c_0 \neq 0$ if and only if x_1 is observed; i.e. $c_1 \neq 0$. As an example of the derivation of the other probabilities consider $P(c_4 = 3, c_{10} = 1)$ when n=12 and m=5. The event $c_4 = 3, c_{10} = 1$ occurs if and only if $w_4 = 1$, $w_5 = 0$, $w_6 = 0$, $w_7 = 1$, $w_{10} = 1$, and $w_{11} = 1$. Thus we must specify the location of four of the observed values and two of the missing values. All arrangements of the remaining three observed and three missing values are assumed to be equally likely, so the required probability is $\binom{6}{3}/\binom{12}{5}$.

In this case the "run" of missing values imputed by x_4 was terminated by the observed value x_7. The form of the probability expression $P(c_i = k, c_j = \ell)$ with $i<j$ depends on whether the run reflected in c_i is terminated by x_j or by some preceding value of $\underline{x}$. Likewise this probability depends on whether the run reflected by c_j is terminated by the end of the sequence of observations or by some value actually observed. This probability also depends on whether $c_i = 0$, $c_j = 0$, or both, and, of course, the distribution of c_0 differs from that of c_i for $i \neq 0$. Table 1 gives the marginal probability distributions $P(c_i = k)$ for selected values of i and k, while table 2 gives the probabilities of $P(c_i = k, c_j = \ell)$ for selected i, j, k, and ℓ. Combinations not given in these tables have probability zero.

It is now a matter of tedious but straightforward algebra to derive the following results, which hold for m=0, 1, 2, ..., n and i=1, 2, 3,...,n:

$$E(c_0) = \frac{1}{(n-m+1)}$$

$$E(c_0^2) = \frac{m(n+m)}{(n-m+1)(n-m+2)}$$

$$E(c_i) = 1 - \frac{\binom{i-1}{n-m}}{\binom{n}{m}}$$

$$E(c_i^2) = \frac{\frac{n+m+1}{n-m+1}\binom{n}{m} + 2(n-m)\binom{i}{m-n+i-1} - (2n+1)\binom{i-1}{m-n+i-1}}{\binom{n}{m}}$$

(See Appendix 3)

Independent Observations - Hot-Deck Procedure

From these results

$$\sum_{i=1}^{n} E(c_i^2) = n + 2m\,\frac{n^2-mn+n-1}{(n-m+1)(n-m+2)} \quad .$$

(See Appendix 4)

Thus, if the elements of $\underline{x}$ are uncorrelated, the variance of $\hat{\mu}$ averaged over all values of $\underline{c}$ is

$$\mathrm{Var}_{HD}(\hat{\mu}) = \frac{\sigma^2}{n}\left[1 + \frac{2m}{n}\,\frac{n^2-nm+n-1}{(n-m+1)(n-m+2)}\right] \quad .$$

It can be shown algebraically that this variance is strictly larger than

$$\frac{\sigma^2}{n}\left(1 + \frac{m}{n-m}\right) \quad ,$$

the variance of the equally-weighted mean under independence, except for the following trivial (or nearly trivial) cases:

1. when m=0, the variances are equal
2. when m=n-1, the hot-deck procedure has a smaller variance (the extra cold-deck value, always available, now has a marked effect on the variance even though the weights are not optimal), and
3. when m=n the equally-weighted mean does not exist, while the hot-deck mean does exist.

(See Appendix 5)

With uncorrelated x_i, $\mathrm{Var}_{HD}(\hat{\mu})$ is approximately

$$\frac{\sigma^2}{n}\left[1 + \frac{2m}{n-m}\right]$$

for large n, whereas

$$\mathrm{Var}_{EW}(\hat{\mu}) \quad \text{is} \quad \frac{\sigma^2}{n}\left[1 + \frac{m}{n-m}\right] \quad .$$

(See Appendix 6)

Thus, averaged over $\underline{c}$ for fixed m, each missing observation (in the case of independence) "costs" about twice as much in increased variance of $\hat{\mu}$ when the hot-deck procedure is used rather than the equal-weights procedure.

As noted, these results are conditional on m but averaged over all values of the random vector $\underline{c}$. The value of $\underline{c}$ will ordinarily be observable, and in the case of independent observations the variance of $\hat{\mu}$ is simply $\sigma^2 \sum_{i=0}^{n} c_i^2$. However, even under the best of circumstances, with nearly equal spacing of the observed and missing observations so that all non-zero values of c_i (including c_0) are [n/(n-m)] or [n/(n-m)] + 1, the variance of the hot-deck procedure will exceed that of the equal-weights procedure except when n/(n-m) is an integer. In the worst case, when one value of c_i equals m+1 and n-m-1 equal unity, the ratio of variances is

$$\frac{(m+1)^2 + (n-m-1)}{n^2/(n-m)} = 1 + \frac{m^2}{n^2}(n-m-1) \quad .$$

Table 3 tabulates for n=15, m=5, and σ^2=1 all possible values of Σc_i^2 and $\mathrm{Var}_{HD}(\hat{\mu})$ conditional on $\underline{c}$. The weighted mean of these estimates is of course equal to the unconditional mean derived above. In this example, $\mathrm{Var}_{HD}(\hat{\mu})$ was always greater than $\mathrm{Var}_{EW}(\hat{\mu})$ and sometimes by a wide margin. The high values for $\mathrm{Var}_{HD}(\hat{\mu})$ in table 3 result from having the missing values in runs or clusters, rather than being evenly spaced. If significant clustering is likely, the hot deck procedure might be undesirable, or one might sort the data to attain more equal spacing of the missing values. Sorting to induce a correlation between successive values of $\underline{x}$ has been suggested above; a sorting scheme to attain both objectives might be difficult to develop, especially if failure to respond ($w_i = 0$) is correlated with the true but unreported value.

Serially Correlated Observations - Hot-Deck Procedure

Assessment of the effect of correlations among the observations requires $E(c_i c_j)$. The values of $E(c_i^2)$ are given above for i=0, 1, 2,...,n. The following results are obtained in a similar way, though the algebra is longer:

$$E(c_0 c_i) = \frac{\left[\binom{n}{m-1} - \binom{i-1}{m-n+i-2} - \binom{n-i+1}{m-i}\right]}{\binom{n}{m-1}} \quad , \qquad i=1, 2, \ldots, n$$

$$E(c_i c_j) = \frac{\left[\binom{n}{m} - \binom{n-j+1}{m-j+1} - \binom{j-1}{m-n+j-1} + \binom{i-1}{m-n+i-1}\right]}{\binom{n}{m}} \quad ; \qquad i \le i < j \le n.$$

(See Appendix 7)

Now assume that $\mathrm{Cov}(x_i, x_j) = \sigma^2 \rho^{|i-j|}$ for i, j=1,2,...,n. One can show that

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(c_i c_j)\rho^{j-i} = \frac{(n+1)(n-m)}{n-m+1}\frac{\rho - \rho^n}{1-\rho}$$

$$- \frac{\rho}{(1-\rho)^2}\left[1 - n\rho^{n-1} + (n-1)\rho^n\right]$$

$$+ \frac{1}{\binom{n}{m}} \frac{\rho^{n-m+1}}{(1-\rho)^{n-m+2}} \sum_{i=n-m+2}^{n} \binom{n}{i}(1-\rho)^i \rho^{n-i}$$

$$- \frac{(n-m)}{\binom{n}{m}} \sum_{i=1}^{n} \binom{n+1-i}{n+1-m} \rho^i \quad .$$

(See Appendix 8)

so that

$$\mathrm{Var}_{HD}(\hat{\mu}) = \frac{\sigma^2}{n^2}\left[n + 2m\,\frac{n^2 - nm + m - 1}{(n-m+1)(n-m+2)}\right.$$

$$+ 2\left\{\frac{(n+1)(n-m)}{n-m+1}\frac{\rho - \rho^n}{1-\rho} - \frac{\rho}{(1-\rho)^2}\left[1 - n\rho^{n-1} + (n-1)\rho^n\right]\right.$$

$$+ \frac{\rho^{n-m+1}}{\binom{n}{m}(1-\rho)^{n-m+2}} \sum_{i=n-m+2}^{n} \binom{n}{i}(1-\rho)^i \rho^{n-i}$$

$$\left.\left. - \frac{(n-m)}{\binom{n}{m}} \sum_{i=1}^{n} \binom{n+1-i}{n+1-m}\rho^i \right\} \right] \quad .$$

The variance of the equal-weights procedure with the same covariance structure was found to be

$$\mathrm{Var}_{EW}(\hat{\mu}) = \frac{\sigma^2}{n^2}\left[\frac{n^2}{n-m} + \frac{2n(n-m+1)}{(n-m)(n-1)}\frac{\rho}{1-\rho}\left(n - \frac{\rho - \rho^n}{1-\rho}\right)\right] \quad .$$

Results to this point are exact, and these formulas are easily usable for relatively small n. For large n, Var_{HD} and Var_{EW} are difficult to compare algebraically in these forms, since the remaining sums cannot be evaluated in closed form without making some approximations. Assume now that $n \to \infty$ and $m/n \to \lambda$. Then for any $\rho \in [-1, 1]$ the term involving the first summation vanishes. (See Appendix 9.) For the second summation, Dr. George Weiss has suggested a method that leads to the following approximation for large n and $0 < \lambda < 1$:

$$\sum_{i=1}^{m} \binom{n+1-i}{n+1-m} \rho^i = \frac{2\rho}{2-\rho} \binom{n}{m-1} .$$

(See Appendix 10)

Using this result,

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(c_i c_j) \rho^{j-i} \doteq n \frac{\rho}{1-\rho} - m \frac{2\rho}{2-\rho} .$$

So that in the limit as $n \to \infty$

$$Var_{HD} \doteq \frac{\sigma^2}{n^2}\left[\frac{n^2+mn}{n-m} + 2\{ n \frac{\rho}{1-\rho} - m \frac{2\rho}{2-\rho} \}\right] .$$

The asymptotic variance of the equal weights procedure with the same covariance structure was found to be

$$Var_{EW} = \frac{\sigma^2}{n^2}\left[\frac{n^2}{n-m} + 2n \frac{\rho}{1-\rho}\right] .$$

The ratio of these two asymptotic variances, with $m = \lambda n$, is

$$\frac{Var_{HD}}{Var_{EW}} = 1 + \frac{\lambda\left(\frac{1}{1-\lambda} - \frac{4\rho}{2-\rho}\right)}{\frac{1}{1-\lambda} + \frac{2\rho}{1-\rho}} ,$$

which, for $\lambda > 0$, is larger than unity provided

$$\lambda > \frac{5\rho-2}{4\rho} .$$

(See Appendix 11)

The ratio is plotted in figure 1 for various values of ρ and λ. Numeric studies show that it attains a minimum of 0.9583 at $\lambda = 0.38$, $\rho = 0.80$. The supremum is 2.0, attained as $\lambda \to 1$ for any value of ρ.

(See Appendix 12)

Minimum Variance Weights

The optimum set of weights $\underline{c}$, given $\underline{w}$ and an arbitrary covariance structure, is an important but unsolved problem. Clearly, weights c must be zero for each missing observation, but how should other weights be assigned so that $\sum_{i=0}^{n} c_i = n$ and $Var(\sum_{i=0}^{n} c_i w_i)$ is minimized? The following results are conditional on the observed $\underline{w}$, not averaged over $\underline{w}$ as in most of the preceding material. A general solution is the smallest root of the determinantal equation $|\sigma^2\underline{\rho} - \underline{c}\underline{I}| = 0$, where $\sigma^2\underline{\rho}$ is the covariance matrix of $\underline{x}$, subject to the constraint that $c_i = 0$ if $w_i = 0$. However, this approach is too general to be useful here.

Let λ be a Lagrange multiplier and let

$$V = [Var(\Sigma c_i x_i) + \lambda(\Sigma c_i - n)]$$

so that for any $c_k \neq 0$

$$\frac{dV}{dc_k} = \frac{d}{dc_k} \sum_{i=0}^{n} \sum_{j=0}^{n} [(c_i c_j \, Cov(x_i x_j) + \lambda(\Sigma c_i - n)]$$

$$= 2 \sum_{i=0}^{n} c_i \, Cov(x_i x_k) + \lambda .$$

Thus, V is minimized when $\sum_{i=0}^{n} c_i Cov(x_i x_j)$ is a constant, $(-n\lambda)/2 = Var(\hat{\mu})$, for all k.

This implies that the optimum values of $\underline{c}$ depend on the covariance structure of $\underline{x}$. Consider the three cases discussed earlier. First, if $\rho_{ij} = 0$ for $i \neq j$ we find that $c_k \sigma^2$ is constant for each non-missing observation. Thus in this case the equal-weights procedure is optimal over all linear alternatives.

For $\rho_{ij} = \rho$, $i \neq j$, the result above becomes

$$c_k\sigma^2 + \sum_{i \neq k} c_i \rho \sigma^2$$

constant, or

$$(1-\rho)c_k + \rho \sum_{i=0}^{n} c_i$$

constant, so that again c_k is constant and the equal weights procedure is optimal.

If $\rho_{ij} = \rho^{|i-j|}$ for all i and j we have

$$\sum_{i=0}^{k-1} c_i \rho^{k-i} + c_k + \sum_{i=k+1}^{n} c_i \rho^{i-k}$$

constant for all k with $c_k = 0$ for each missing value. This expression is difficult to work with directly, but when ρ is small it may be sufficient to consider only those terms in the sums such that $|i-k|$ is small, say $|i-k| \leq j$

intended sample. Second, imputation of specific values by the hot deck method permits the easy estimation of various cells in cross-tabulations, while this may be more difficult (and provide a different answer) with the equal-weights procedure. This, too, needs exploration.

Other problems that remain unexplored are the effects of clustering of missing observations, the effects of sampling from finite populations, extensions of the analysis to problems other than the estimation of item means (e.g. the estimation of variances) and the use of non-linear functions of observed values to impute missing values. We have assumed here that "sampling weights" are all equal; when sample elements have varying weights in the analysis, as in sampling proportionale to size, our assumptions about the structure of $\underline{c}$ are no longer valid. This also needs exploration.

We believe that it is particularly important to undertake theoretical studies of the effect of hot-deck procedures on bias, since those procedures are commonly justified in terms of reduction of bias rather than control of variance, convenience, or other assumed virtues. To some extent, at least, sorting records to maximize serial correlations might tend to trade reduced bias for increased variance in both equal-weights and hot deck procedures, since missing values would tend to be clustered in parts of the data set where observed values are considerably above or below the mean. Thus the various possible sequences of location for missing values would no longer be equally likely, and our assumptions about $\underline{c}$ would change. To the extent that bias would be traded for increased correlation (and hence increased variance of $\hat{\mu}$), the bias-reducing properties of the hot-deck procedure are already accounted for in the present analysis, but the matter needs further study.

If we ignore the cold deck value and special problems at either end of the sequence of sample elements, and if we assume negligible probability that more than one observation will be **missing in any string of 2j+1 sample elements, we can use**

$$\sum_{i=k-j}^{k-1} c_i \rho^{k-i} + c_k + \sum_{i=k+1}^{k+j} c_i \rho^{i-k} \simeq c_k + \sum_{i=1}^{j} (c_{k+i} + c_{k-1}) \rho^{i-k}$$

$$= \text{Constant} .$$

This implies, for sufficiently small ρ, that we impute the observed value nearest the missing value (not just the nearest preceding value), except that if the missing value is equidistant from the nearest preceding and following values, we impute their mean. This seems intuitively reasonable. More generally, each missing observation is imputed by a weighted sum of nearby values. We have not worked out the variance of this modification of the hot-deck procedure.

Conclusions

In the three cases examined here, the standard hot-deck method is inferior to the equal-weights method for the limited purpose of estimating item means.

Two important considerations may still improve the value of the hot-deck procedure relative to the equal-weights procedure. First, we have assumed to this point that $\underline{w}$ is independent of $\underline{x}$, implying that whether an observation is missing is independent of its true value. Under this assumption, both estimates are unbiased, while if it is violated both are in general biased, but to different degrees, and perhaps even in different directions. This matter should be explored to determine how each procedure performs when missing values are not a random subset of the