MISSING DATA PROCEDURES:

A COMPARATIVE STUDY

Sampling Studies Section
Sample Surveys Research Branch
Statistical Reporting Service
United States Department of Agriculture
Washington, D.C.

August 1976
SF 76-02

MISSING DATA PROCEDURES:

A COMPARATIVE STUDY

by

BARRY L. FORD

Sampling Studies Section
Sample Surveys Research Branch
Statistical Reporting Service
United States Department of Agriculture
Washington, D.C.

August 1976

TABLE OF CONTENTS

PURPOSE

The purpose of this paper is to discuss an investigation of the missing data problem in a list frame survey which has a simple stratified design. A description of six missing data procedures is given, and problems in their application are discussed. The main thrust of this report is a simulation experiment with these missing data procedures. The data are from an agricultural survey by the Statistical Reporting Service (SRS) of the United States Department of Agriculture (USDA).

INTRODUCTION

In the area of survey design, the missing data problem is one of increasing concern. Non-response rates of 10% are not unusual for SRS surveys, and there is a fear that these rates may increase.

Why worry about missing data? If there is little difference between the missing data and the reported data in a simple stratified design, the only consequence of missing data is the reduction in sample size. This reduction can easily be offset by increasing the initial sample size. However, in many cases it is probable that the missing data and the reported data are not alike.

A difference between missing and reported data leads to biases in the survey estimates. The size of these biases depends on:

1: the magnitude of the difference between the missing and reported data

2: the percentage of nonresponse .

The causes of missing data are complex and varied, but the emphasis in any survey should be on eliminating or minimizing the likelihood of missing data before the survey starts. Procedures to estimate for missing data are a stopgap measure -- they are techniques to use after the survey is over when no other alternative is possible. Obviously, no procedure can be as good as not having any missing data. Furthermore, when the percentage of missing data is extremely high, there is probably no procedure that can estimate the missing

data efficiently enough to make the survey worthwhile. With moderate and low missing data rates, perhaps some missing data procedures can minimize the bias to a tolerable level.

The six missing data procedures discussed in this investigation are the double sampling ratio procedure, the double sampling regression procedure, and four variations of the hot deck procedure. Some general advantages and disadvantages of each one are outlined.

The Double Sampling Ratio Procedure

Often there is an auxiliary variable associated with each sampling unit. This auxiliary variable may be one that is used to stratify the population, an observed variable, or any other additional variable that can be obtained for the whole sample. There should also be a reasonable correlation between the primary variable (the variable of interest) and the auxiliary variable.

In a missing data context the first sample is the selected sample, including missing and reported data. The second sample is only the reported data. The ratio estimator and its approximate variance for a simple random sample (1, pg. 340) are:

(I) $$\bar{y}_{Ratio} = \frac{\bar{y}}{\bar{x}}\,\bar{x}'$$

(II) $$VAR(\bar{y}_{Ratio}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[S_y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)(2R\,S_y S_x - R^2 S_x^2)\right]$$

where:

$\bar{x}'$ = average of the auxiliary variable over the whole sample

$\bar{x}$ = average of the auxiliary variable over the part of the sample that reported data

$\bar{y}$ = average of the primary variable over the part of the sample that reported data

$\bar{X}$ = the average of the auxiliary variable over the whole population

$\bar{Y}$ = the average of the primary variable over the whole population

$R = \frac{\bar{Y}}{\bar{X}}$

S_x^2 = the variance of the auxiliary variable

S_y^2 = the variance of the primary variable

ρ = the correlation between x and y

n' = size of the entire sample

n = size of the sample that reported data

N = size of the population

(Note that the variance was multiplied by the finite population correction factor).

Although the double sampling ratio estimator is almost always a biased estimator, it is easy to compute even for complex samples. In this report the design is a simple stratified sample so the above formulas are applied in each stratum. Usually S_y^2, S_x^2, ρ, and R are unknown, but their corresponding sample estimates can be substituted into the previous two equations (I and II). As Cochran points out (1, pg. 341), the resulting estimate of variance is not unbiased but appears to be a good approximation.

This ratio estimator makes two assumptions:

1. the initial sample is a random sample

2. the missing data comprise a *random* subsample of the initial sample.

This second assumption is probably violated in most surveys; to what degree it is violated depends of course, on the particular situation. One hopes that the ratio estimate and its variance are fairly insensitive to a violation of the second assumption.

In essence the ratio estimator is a linear regression estimator with the intercept assumed to be zero. If the population does not follow the assumption of a linear model, then the ratio estimator (or any regression estimator) becomes a biased estimator. Researchers rarely accept the linear population

model as completely realistic, but approximate analytical results and empirical studies show the bias is usually small (3, pg. 25-25; 8, pg. 208-209; 9).

One should remember that in a stratified design there also exists a combined ratio estimator. This estimator is used when the ratio $R = \frac{\bar{Y}}{\bar{X}}$ is equal in all strata and the sample size is small in each stratum. For the data in this study the idea that the ratios in all strata are equal is believed to be false. Thus, a separate ratio estimator is used in each stratum. However, the separate ratio estimator has an inherent danger of accumulating a serious bias across all strata. This accumulation is more likely to be serious when the stratum biases are in the same direction (1, pg. 168-173).

The Double Sampling Regression Procedure

The double sampling regression procedure is also quite commonly used. Like the ratio procedure one has an auxiliary variable in addition to the primary variable. The formulas are (1, pg. 336-339):

(III) $$\bar{y}_{Reg} = \bar{y} + b\,(\bar{x}' - \bar{x})$$

(IV) $$VAR\,(\bar{y}_{Reg}) = \frac{S_y^2\,(1 - \rho^2)}{n} + \frac{\rho^2 S_y^2}{n'}$$

We will estimate VAR $(\bar{y}_{Reg})$ with:

$$var\,(\bar{y}_{Reg}) = \frac{s_{y.x}^2}{n} + \frac{s_y^2 - s_{y.x}^2}{n'}$$

Adjusting var $(\bar{y}_{Reg})$ by a finite population correction factor of $1 - \frac{n}{N}$, one obtains:

(V) $$var\,(\bar{y}_{Reg}) = \left(1 - \frac{n}{N}\right)\left[\frac{s_{y.x}^2}{n} + \frac{s_y^2 - s_{y.x}^2}{n'}\right]$$

as an estimate of the variance of $\bar{y}_{Reg}$ in a finite population where:

$\bar{x}'$, S_y^2, ρ^2, n', n, N are the same as for the ratio estimator

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

$$s^2_{y.x} = \frac{1}{n-2} \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] .$$

As noted for the ratio estimator, the assumption of a linear population model may be false. The estimators may then be biased, but one empirical study (3, pg. 22-25) lends support to the conjecture that these biases are small. Also, in the stratified data of this study a separate regression estimator is used as opposed to a combined regression estimator. Therefore, one again has the danger of accumulating a large bias across all strata, especially when each stratum bias is in the same direction (1, pg. 200-202).

Hot Deck Procedure -- Random Substitution

The hot deck is probably the most common missing data procedure in use at the present time, especially in complex surveys. The Bureau of the Census, the Statistical Reporting Service, Statistics Canada, and many other agencies currently employ this missing data procedure. In spite of this wide use little testing or theoretical analysis on the impact of the hot deck procedure has been published (9). This situation is really not too surprising because although the hot deck procedure is intuitively satisfying and extremely flexible, its flexibility and lack of a strong theoretical development deter anything but broad generalizations of its effects.

A basic outline of the hot deck procedure is:

1: Separate the sample into I classes based on k variables.

2: If an item is missing in a certain class, then randomly select a reported item from the same class.

3: Substitute the selected item for the missing item.

4: Compute sample estimates as if there are no missing values.

One consequence of this procedure is that the estimated variances of the sample mean tend to be biased below their actual values. This tendency is stronger when the missing data and reported data are different in distribution. Step 4 above allows one to use a sample size that includes the number of missing values. Thus, the loss of information due to missing data is not reflected in the sampling errors. For example, suppose two surveys cover the same population and have the same sample size. Furthermore, one survey has 30% missing data, and the other survey has no missing data. After applying the hot deck procedure, the errors of the estimates of these two surveys would probably be about equal. Even in the case, where the missing and reported data have similar distributions, *the standard errors from the survey that used the hot deck procedure would reflect the fact that 30% of the information is missing.*

One should also note that the *sample elements are no longer independent.* The hot deck procedure is essentially a duplicating process with reported values substituting for missing values. The covariance that results from this duplication is ignored in the hot deck procedure. Ignoring this covarinace can be a serious error.

Probably the greatest attraction of the hot deck procedure is its operational simplicity. The classification of the data items into I classes is an extremely adaptable method. The classification variables may be cardinal, ordinal, categorical, etc. In fact, the whole classification method may vary from the subjective to the mathematical rigorous. In addition, many complex surveys will not use the hot deck procedure because of the pressure to retain the planned sample design (eg. self-weighting designs, survey designs using balanced repeated replications, etc.).

The looseness of the classification method has tended also to obstruct theoretical evaluations of the hot deck procedure and thus to impede any theoretical comparisons between it and other missing data procedures. Some general theory of the way the hot deck procedure reduces the bias due to missing data is in the Appendix A.

Hot Deck Procedure -- "Closest" Substitution

One possible alternative to the random substitution of the hot deck procedure is to substitute the "closest" reported item for each of the missing items. With one auxiliary variable, the "closest" value to a missing item is simply the value for which the absolute difference between the auxiliary variable of the missing item and the auxiliary variable of the reported item is minimized. In the case of the ties for the "closest" auxiliary value a random selection of one of the tied values is made.

This procedure should have the same effect as assigning the population to many strata and selecting a few units from each stratum (since the stratification is based on the auxiliary variable). Thus, suppositions that the hot deck method improves with more narrowly defined strata can be examined with the results of the "closest" procedure.

Given a good range coverage, this procedure is fairly robust to very curved relationships between the auxiliary and primary variables. The data in this investigation is not curved enough to reveal this robust property of the "closest" procedure.

Hot Deck Procedure -- "Two Closest" Substitution

This procedure is another variation of the hot deck procedure. Instead of substituting the "closest" reported item for each missing item, one substitutes the average of the "closest" value whose auxiliary value is smaller than the reported item and the "closest" value whose value is larger than the reported item.

Hot Deck Procedure -- "Class" Mean Substitution

This last variation of the hot deck procedure substitutes the average of the reported units in a class for each missing unit in that class. It is the simplest of the procedures presented in this paper.

THE SIMULATION EXPERIMENT

Why Use Simulation?

The need for simulation in this investigation is to compare the estimated variances of the estimated means. Possibly one might be able to compare how differences in the missing and reported data theoretically affect the estimated means using these six missing data procedures. However, the problem of analytically comparing the estimated variances of the estimated means is unreasonable. The fact that some assumptions fail in each procedure complicates the analytical work.

For example, one should recall the estimated mean of the hot deck procedure, $\bar{x}_{HD}$. Assuming there are differences in the missing and reported data, one can not explicitly write the expected value of the estimated variance of $\bar{x}_{HD}$, $E[\hat{Var}\ (\bar{x}_{HD})]$. In fact, it is not known if $E[\hat{Var}\ (\bar{x}_{HD})] = Var\ (\bar{x}_{HD})$, and the author strongly doubts that it does. However, this paper will provide no evidence to support that supposition because the structure of the simulation of this experiment does not allow an estimate of $Var\ (\bar{x}_{HD})$ but does allow an estimate of $E[\hat{Var}\ (\bar{x}_{HD})]$. If $E[\hat{Var}\ (\bar{x}_{HD})] \neq Var\ (\bar{x}_{HD})$, then there is quite a weakness in the hot deck procedure. The costs of a simulation experiment providing this type of evidence would be much greater than the simulation actually used. This investigation contents itself with comparing the estimated variance of the estimated mean for each procedure with the estimated variance

if the sample had no missing data. These comparisons will serve the purpose of revealing certain key qualities of each procedure.

As noted before the double sampling ratio and regression procedures also have variance estimates that involve assumptions and approximations that may be tenuous. For example, the assumption of a linear model is usually invalid in the regression and ratio procedures, and the ratio procedure simply uses substitution as an approximation to variance estimation. On the basis of two important studies (3;8) and practical experience one does not expect these biases in the variance estimates to be substantial for large samples. However, the comparisons among the procedures may be sensitive enough that these biases would be large enough to affect the comparisons.

Analysis

The primary point in the comparison of these procedures will be the minimization of the biases caused by missing data in the estimated means. Secondary importance is given to the comparisons of the estimated variances of the estimated means. Details of the experimental design used in this study are in the Appendix B.

The data in this simulation are from a simple stratified design, i.e. the sample within each stratum is a simple random sample independent of the samples in other strata. An auxiliary variable is used to assign the population to nine strata from which the initial samples are selected. The sample sizes are shown in Table 1. One should note that Strata 1 and 2 both have a value of zero in the auxiliary variable and are separated on the basis of a second nominal variable.

Table 1 shows the correlations between the auxiliary variable and the primary variable. As one can see, the correlations in the first two strata are zero, by virtue of the fact that the auxiliary variable only has one value, zero, in these two strata. Furthermore, the correlation in stratum 9 is essentially zero.

One may think that with larger correlations between the primary and auxiliary variable the estimates from the regression procedure would improve dramatically compared to the other procedures. However, the data in this study prevent evidence for or against this hypothesis. Surely, larger correlations would improve estimates resulting from all the procedures. Whether this improvement is equal for all the procedures is the question which can not be answered in this report.

Table 1: Stratification and sample sizes of the data used in the simulation experiment.

Stratum	Auxiliary Variable	Correlation Between Auxiliary and Primary variables	Population Size	Sample Size
1	0	0.00	29,360	257
2	0	0.00	26,343	184
3	1-199	0.35	25,618	238
4	200-349	0.27	16,564	206
5	350-599	0.22	13,600	232
6	600-999	0.32	6,368	180
7	1000-1999	0.26	2,275	106
8	2000-4999	0.43	448	98
9	5000+	0.03	55	21

For the hot deck classification each stratum was divided into four classes of approximately the same span in the auxiliary variable. For example, Stratum 3 has sampling units with the value of the auxiliary variable ranging from 1-199. Thus, the four hot deck classes for Stratum 1 allow the auxiliary variable to range from 1-49, 50-99, 100-149, and 150-199. Strata 8 and 9 have three and two classes respectively because one does not wish to make the number of sample units in a certain class very small since the computer program may simulate all the units in a class as missing.

Results

From the analysis of variance (see Appendix B), the six missing data procedures do not yield significantly different estimates of the mean. The test statistic:

$$F = \frac{MS_T}{MS_{TxPlots\ [Within\ AxB]}} = \frac{.929}{1.023} = 0.91$$

has 5 and 855 degrees of freedom, and the significance of the statistics is a little over 50%.

One should note that Tukey's multiple comparison test could locate differences of 0.3% between any two of the estimated means of the six procedures, and that is certainly accurate enough for this practical survey application.

The average improvement in the estimated mean using any of the six missing data procedures is shown in Table 2. The missing data procedures are compared for different levels of bias (B), caused by using only the reported data to estimate the mean of the population. Using only the reported data is in fact, the current SRS procedure. Missing reports are omitted, and the expansion factors are simply adjusted. Table 3 displays the percentage reduction in bias from using any of the six missing data procedures. Since there is no significant difference in the estimated means, average improvements over the six procedures are shown in Table 3 for the six cases where there is a bias in the reported data.

Table 2: The difference between the estimated means from a simulation of the missing data and 94,889, the estimated mean if the sample has no missing data

Bias*	Estimated Means Minus 94,889					
	Double Sampling Ratio Procedure	Double Sampling Regression Procedure	Hot Deck Procedures			
			Random Substitution	"Closest"	"Two Closest"	"Class" Mean
0.0 (5% missing items)	-0.090	-0.087	-0.163	0.251	0.179	-0.199
0.0 (10% missing items)	-0.474	-0.469	-0.131	0.075	-0.053	-0.604
0.0 (20% missing items)	0.382	0.379	0.788	0.931	0.754	0.008
-1.9	-1.364	-1.354	-1.405	-1.378	-1.446	-1.445
-2.9	-2.637	-2.626	-2.482	-2.839	-2.682	-2.726
-4.0	-3.374	-3.495	-3.793	-3.744	-3.750	-3.653
-6.2	-5.377	-5.335	-5.504	-5.719	-5.652	-5.482
-9.0	-7.381	-7.347	-7.608	-7.887	-8.032	-7.619
-14.0	-11.205	-11.174	-11.412	-11.124	-11.300	-11.418
Average Over All Levels of Bias	-3.502	-3.501	-3.523	-3.493	-3.554	-3.682

*These levels of bias are determined by the percentage of missing items and the difference in the means of of the reported data and the missing data. For example, a bias level of -1.9 is due to 5% of the data missing and to a difference of -35 between averages of the reported and missing data. Appendix B contains further details.

Table 3: Average improvement in the estimated mean from the simulation of six missing data procedures

B= Bias (Mean of the reported data minus the true sample mean)	A=Average estimated mean of the six missing data procedures minus the true sample mean	$\frac{B-A}{B} \cdot 100\%$
-1.9	-1.40	26%
-2.9	-2.66	8%
-4.0	-3.63	9%
-6.2	-5.51	11%
-9.0	-7.65	15%
-14.0	-11.27	20%

Since there is no significant difference in the estimated means among the six procedures, the focus of interest becomes the estimated variances of the estimated means. The analysis of variance for the six missing data procedures with the estimated variance as the dependent variable is shown in Appendix B. Obviously, there is a significant difference among the estimated variances because the test statistic is so large:

$$\frac{MS_T}{MS_{TxPlots\ [Within\ AxB]}} = \frac{162.245}{0.391} = 414.95$$

Performing Duncan's multiple comparison test at a 95% significance level on the estimated variances separates the procedures into the following groups:

{ t_1: double sampling ratio procedure
{ t_2: double sampling regression procedure

{ t_3: hot deck procedure with random substitution
{ t_4: "closest" procedure
{ t_5: "two closest" procedure

{ t_6: "class" mean procedure.

Perhaps it is more revealing to examine the estimated variances at each of the nine levels of bias used in simulation. Table 5 shows the estimated variance using each procedure minus 9.685, the estimated variance if the sample has no missing data. Figure 1 is a graphical display of the results.

The usual criterion for judging the variances of the estimated means resulting from the missing data procedures is that the smallest variance is the best. However, a procedure may result in a small estimated variance simply because of a large negative bias. By comparing the estimated variances resulting from the procedures with the estimated variance if the sample has no missing data, one can judge if there are any large negative biases in the estimated variances. For example, if the estimated variance resulting from a missing data procedure is 7.20, then obviously there is a large negative bias in estimating the variance since the estimated variance of the sample has no missing data is 9.685.

One first notices that in Table 5, as in all the results, there is little difference between the estimated variances for regression and ratio estimates (procedures 1 and 2). One then notes that in the first three cases there is zero bias, and the hot deck estimator with random substitution (procedure 3) yields variances close to 9.685. The ratio and regression procedures have larger differences because they depend on the weak correlations of the primary and auxiliary variable. The "closest" procedure, the "two closest" procedure, and the "class" mean procedure (4, 5 and 6) have negative values. The negative values indicate *that their estimated variances are even smaller than the estimated variance if the sample has no missing data*,

Table 4: The difference between the estimated variance of the estimated mean resulting from a simulation of the missing data and 9.685, the estimated variance of the estimated mean if the sample has no missing data.

B=Bias	Estimated Variance of the Estimated Mean Minus 9.685					
	Double Sampling Ratio Procedure	Double Sampling Regression Procedure	Hot Deck Procedures			
			Random Substitution	"Closest"	"Two Closest"	"Class Mean
0.0 (5% missing items)	0.556	0.556	0.126	-0.024	-0.045	-0.528
0.0 (10% missing items)	0.885	0.883	0.055	-0.701	-0.295	-1.238
0.0 (20% missing items)	2.478	2.478	0.580	-0.785	-0.415	-1.971
-1.9	0.263	0.263	-0.095	-0.416	-0.454	-0.781
-2.9	-0.223	-0.224	-0.654	-0.970	-0.906	-1.204
-4.0	0.355	0.353	-0.884	-0.973	-1.094	-1.685
-6.2	0.096	0.094	-0.950	-1.277	-1.385	-1.872
-9.0	1.087	1.084	-1.205	-1.434	-1.700	-2.928
-14.0	-0.086	-0.090	-1.928	-2.579	-2.750	-3.571
Average Over All Levels of Bias	0.601	0.600	-0.551	-1.018	-1.005	-1.753

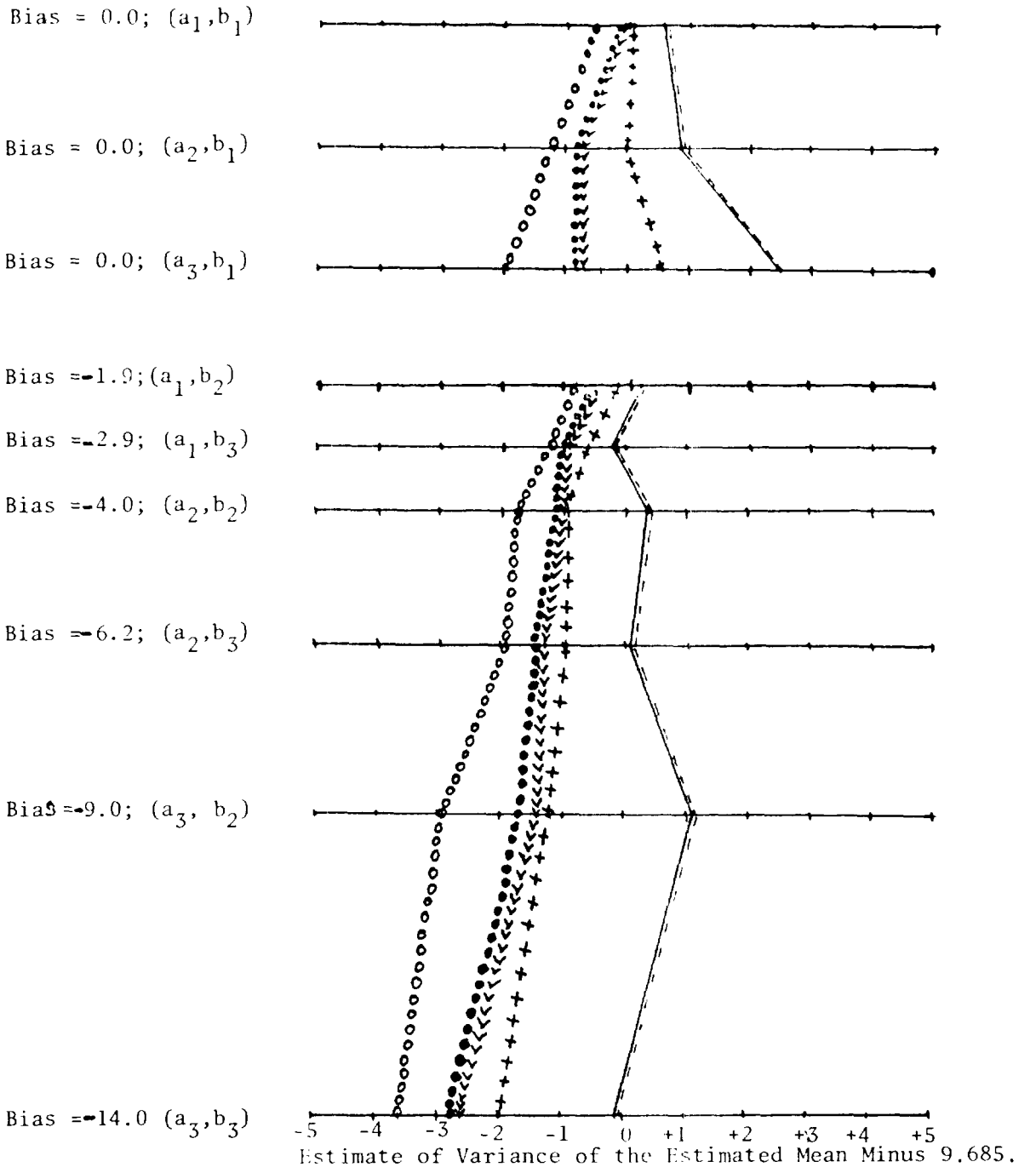

Figure 1: Graph of the estimated variance of the estimated mean of the six missing data procedures minus 9.685 the estimated variance of the estimated mean if the sample had no missing items.

CONCLUSIONS

This investigation offers a comparison of six missing data procedures as applied to a specific data set taken from a SRS list frame survey. This data set is from a simple stratified sample with a large sample size in each stratum. The following procedures are applied to this data set:

1: the double sampling ratio procedure

2: the double sampling regression procedure

3: a hot deck procedure in which a randomly selected reported item was substituted for each missing item

4: the "closest" procedure in which the closest reported item was substituted for each missing item

5: the "two closest" procedure in which the average of the two closest reported items was substituted for each missing item

6: the "class" mean procedure in which the "class" mean was substituted for each missing item.

The most important aspect in comparing these missing data procedures is to protect against biases in the estimated means (or totals). An analysis of variance shows no significant differences among the estimated means which result in using these procedures. *All the procedures reduce the relative bias that results from accepting the mean of the reported data as an estimate of the population mean.* This relative bias is a result of the non-response rate and a difference in the variable of interest (e.g. total hogs, total hogs, total cattle) between the respondents and non-respondents. The reduction in relative bias averages 15% and varies from 8% to 26%. Considering the low correlations between the auxiliary and primary variables, this reduction is reasonable. Much larger reductions in bias could be obtained if a variable could be found with a high correlation with the variable of interest and if that variable could easily be obtained for the entire sample. Of course,

sample. Of course, there is a great deal of room for improvement over the 15% reduction, and later research should examine other procedures (including multivariate ones) to evaluate their efficiency.

An important, though secondary, concern is the estimated variances of the estimated means. All of these estimated variances except those from the ratio and regression procedures are underestimates of the true variances because they are generally less than the estimated variance that result with no missing data in the sample. Furthermore, the degree of underestimation *increases* as the relative bias increases. This part of the investigation clearly reveals why all of the hot deck procedures (random, "closest", "two closest", and "class" mean substitution) may be undesireable. It does not seem reasonable to use one of the hot deck procedures when it does not reduce the bias of the estimated mean any more than the ratio or regression procedure, and yet the variance of the mean is greatly underestimated. Probably, there is an underestimation of variance in the ratio and regression procedure, but these results show that it is not nearly as large as in the hot deck procedures.

The final result of this investigation is a recommendation of the ratio or regression procedures (the effects of these two procedures being indistinguishable). These two procedures have been more theoretically explored than the other procedures. The estimated variances of the estimated means from the ratio or regression procedure reflect better than the other procedures the true quality of the data. Finally, the ratio or regression procedure can be easily implemented into the SES estimates by simply adding the sample control data to the computer data tape.

BIBLIOGRAPHY

1. Cochran, William G. Sampling Techniques, John Wiley and Sons, Inc. 1963.

2. Dear, R.E. "A Principle Component Missing Data Method for Multiple Regression MOdels" Systems Development Corporation, Santa Monica, California, SP - 86, 1959.

3. Frankel, Martain R. Inference From Survey Samples: An Empirical Investigation, Institute for Social Research, University of Michigan, 1971.

4. Hartley, H.O. and Hocking, R.R. "The Analysis of Incomplete Data", Biometrics, Volume 27, pages 783-824, 1971

5. Hocking, R.R. and Smith, W.B. "Estimation of Parameters in the Multiple Normal Distribution with Missing Observations", Journal of the American Statistical Association, Volume 63, pages 159-173, 1968.

6. Kirk, Roger E. Experimental Design: Procedures for the Behavioral Sciences, Brooks Cole Publishing Company, 1968.

7. Kish, Leslie Survey Sampling, John Wiley and Sons, Inc. 1965.

8. Kish, Leslie; Namboodiri, N.K.; and Pillai, R.K. "The Ratio Bias in Surveys", Journal of the American Statistical Association, Volume 57, pages 863-876. 1962.

9. Rockwell, Richard C. "An Investigation of Imputation and Differential Quality of Data in the 1970 Census", Journal of the American Statistical Association, Volume 70, pages 39-42, 1975.

APPENDIX A

The hot deck does have some simple qualities to recommend it. For example, let $E(\bar{x} - \mu) = B$ be the bias associated with nonresponse when estimating the population mean, μ, with the mean of a simple random sample. To estimate μ using the hot deck procedure one divides the sample data into I classes. Let $E(\bar{x}_i - \mu_i)$ be the bias in class i, i = 1, 2, ..., I. If p_i^* is the proportion of the population in class i, then the bias B_{HD}, associated with the estimated mean, $\bar{x}_{HD}$, of the sample data after applying the hot deck procedure is simply:

$$B_{HD} = E(\bar{x}_{HD} - \mu) = \sum_{i=1}^{I} p_i^* B_i .$$

To prove this equation one notes:

$$E[\bar{x}_{HD}] = E\left[\sum_{i=1}^{I} p_i \bar{x}_i\right] = E_{n_i}\left[E\left\{\sum_{i=1}^{I} p_i \bar{x}_i \mid n_i\right\}\right]$$

where n_i is the number of sample units that fell in class i, n is the total sample size, $p_i = \frac{n_i}{n}$, and $\bar{x}_i$ is the sample average for class i. The expected value inside the braces is over fixed n_i, and the expected value outside the braces is then over all possible values of n_i. Obviously,

$$E_{n_i}\left[E\left\{\sum_{i=1}^{I} p_i \bar{x}_i \mid n_i\right\}\right] = E_{n_i}\left[\sum_{i=1}^{I} p_i E(\bar{x}_i)\right]$$

$$= \sum_{i=1}^{I} p_i^* E(\bar{x}_i) .$$

Since $B_{HD} = E(\bar{x}_{HD}) - \mu$, then

$$B_{HD} = \sum_{i=1}^{I} p_i^* E(x_i) - \mu$$

$$= \sum_{i=1}^{I} p_i^* E(\bar{x}_i - \mu_i)$$

$$= \sum_{i=1}^{I} p_i^* B_i .$$

In spite of the fact that $\bar{X}_i$ and n_i are not independent, they are uncorrelated. Now, if $|B_i| < |B|$, for each i then:

$$(I) \qquad |B_{HD}| = \left|\sum_{i=1}^{I} p_i^* B_i\right| < \sum_{i=1}^{I} p_i^* |B_i| < \sum_{i=1}^{I} p_i^* |B| = |B|.$$

Thus, one can see that the bias using the hot deck procedure is less than the bias caused by omission of missing data *on the condition that* $|B_i| < |B|$ for each i. This condition should hold in most cases, but there is no guarantee because it is a function of the quality of the classification method. A good classification method should decrease the absolute value of the bias below $|B|$ in each of the I classes. However, the hot deck procedure allows <u>any</u> classification. The goodness of the classification process is left to the integrity of the statistician.

APPENDIX B

An important aspect of this study is the fact that the comparisons are based *on an experimental design where each observation is the result of a simulation of a procedure*. This situation is quite different from a simulation study where there may be a thousand or more simulations in order to narrow the confidence interval of an estimate almost to a point. Because of constraints on money, there are only a total of 180 simulations on each procedure. The design of the experiment partly compensates for the resulting loss of accuracy. In fact, the experimental design preserves enough accuracy to distinguish *reasonably* between different procedures.

The experimental design of this study is a split plot design with two block effects between the plots and one treatment effect within each plot. The two block effects and their levels are:

A: the fixed effect of the percentage of missing items [3 levels]

a_1 : 5% missing items

a_2 : 10% missing items

a_3 : 20% missing items

B: the fixed effect of the difference between the total population mean and the mean of the missing items [3 levels]

b_1 : $\mu - \mu_2 = 0$

b_2 : $\mu - \mu_2 = -35$

b_3 : $\mu - \mu_2 = -55$

where μ is the population mean and μ_2 is the mean of the sample units that would be missing.

The treatment effect and its levels are:

T: the fixed effect of a procedure on the estimate of the mean [6 levels]

t_1 : double sampling ratio procedure

t_2 : double sampling regression procedure

t_3 : hot deck procedure (random substitution)

t_4 : "closest" procedure

t_5 : "two closest" procedure

t_6 : "class" mean procedure.

Finally, a plot is one computer simulation of the missing items. Each simulation consists of randomly deleting items from each stratum of a simple stratified sample. Each procedure is then applied to the remaining items. There are n = 20 plots for each level of A x B. A plot then has six observations made on it -- an observation being an estimate using a certain procedure.

The two block effects, A and B, are functionally related in terms of the bias, *B*, associated with estimating the mean with *only* the reported items. In the case of a simple random sample where:

p = the percentage of the population that does not have missing data

q = the percentage of the population that has missing data,

one knows:

$$\mu = p\mu_1 + q\mu_2$$

where, μ_1 = the mean of the reported items. Thus,

$$B = \mu_1 - \mu = \frac{q}{p}(\mu - \mu_2), \text{ where } p \neq 0.$$

Under the different level combinations of the block effect the biases and the relative biases, $RB = \frac{B}{\mu}$, are exhibited in Table 5.

Table 5: The bias, *B*, and relative bias, *RB*, caused by not estimating the missing data items in a simple random sample

Level of A x B	*B*=Bias	*RB*=Relative Bias
(a_1,b_1)	0.0	0.0%
(a_2,b_1)	0.0	0.0%
(a_3,b_1)	0.0	0.0%
(a_1,b_2)	-1.9	-2.0%
(a_2,b_2)	-4.0	-4.2%

(a_3, b_2)	-9.0	-4.2%
(a_1, b_3)	-2.9	-3.1%
(a_2, b_3)	-6.2	-6.5%
(a_3, b_3)	-14.0	-14.8%

One can see in Table 5 that there is a good range in the degree of the bias. Although this study uses a simple stratified sample, rather than a simple random sample, the biases in Table 6 apply because it is assumed that the percentage of missing items is the same from stratum to stratum. In other words, in a stratified design:

$$\sum_{\text{All strata } i} \frac{q_i}{p_i} \rho_i (\mu_{1_i} - \mu_{2_i}) = \frac{q}{p} (\mu_1 - \mu_2)$$

because $p_i = p$ and thus $q_i = q$ for each i (ρ_i = the proportion of the total population in stratum i).

The model for the split plot design of this study is (adapted from 7, pg. 284):

$$y_{ijkm} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \pi_{(ij)m} + \tau_k + (\alpha\tau)_{ik} + (\alpha\beta\tau)_{ijk} + \tau\pi_{(ij)km} + \varepsilon_{(ijk)m}$$

i = 1, 2, 3

j = 1, 2, 3

k = 1, 2, ..., 6

m = 1, ..., 20

α_1 = the effect of the ith level of A

β_j = the effect of the jth level of B

$(\alpha\beta)_{ij}$ = the effect of the interaction of the ith level of A and jth level of B

$\pi_{(ij)m}$ = the error between plots within each level of A x B

$\varepsilon_{(ijk)m}$ = the error between observations within each plot

τ_k = the effect of the kth level of T

$(\alpha\tau)_{ik}$ = the effect of the interaction of the ith level of A and kth level of T

$(\beta\tau)_{jk}$ = the effect of the interaction of the jth levle of B and the kth level of T

$(\alpha\beta\tau)_{ijk}$ = the effect of the interaction of the ith level of A, jth level of B, and the kth level of T

$\tau\pi_{(ijk)m}$ = the effect of the interaction of the kth level of T and the mth plot within each level of A x B

and $\varepsilon_{(ijk)m} \sim N(0,\sigma^2_\varepsilon)$.

Then Table 6 (derived from 6, page 287 and 293) exhibits the appropriate analysis of variance table given that a = 3, b = 3, t = 6, and n = 20.

(Sums of squares and mean square formulas can be found in 7, page 285-286).

Table 6: Analysis of variance table for a split plot desing with two factors, A and B, between the plots and one factor, T, within each plot.

Source of Variation	Degrees of Freedom
Between Plots	nab-1 = 179
A	a-1 = 2
B	b-1 = 2
A x B	(a-1)(b-1) = 4
Between Plots [Within A x B]	ab(n-1) = 171
Within plots	nab(t-1) = 900
T	t-1 = 5
T x A	(t-1)(a-1) = 10
T x B	(t-1)(b-1) = 10
T x A x B	(t-1)(a-1)(b-1) = 20

T x Plots [Within A x B]	ab(n-1)(t-1) = 855
Total	nabt - 1 = 1079

The test of interest concerns:

H_o: no difference in the effect of the procedures

H_a: a difference in the effect of the procedures .

Therefore, the test statistic is:

$$F = \frac{\text{Mean Square}_T}{\text{Mean Square}_{T \times \text{Plots [Within A x B]}}}.$$

F follows an *F*-distribution with 5 and 855 degrees of freedom.

The analyses of variance with the estimated mean as the dependent variable and the estimated variance of the estimated mean as the dependent variable are in Tables 7 and 8 respectively.

Table 7: Analysis of variance on six missing data procedures in which the dependent variable is the estimated mean.

Source	Degrees of Freedom	Mean Square
Between Plots	179	93.055
A	2	2087.368
B	2	4005.478
AxB	4	721.686
Between Plots Within AxB	171	9.266
Within Plot	900	1.014
T	5	0.929
TxA	10	0.157
TxB	10	2.321
TxAxB	20	0.451
TxPlots [Within AxB]	855	1.023
Total	1079	

Table 8: Analysis of variance of the estimated variance of the estimated mean from the simulation of the six missing data procedure.

Source	Degrees of Freedom	Mean Square
Between Plots	179	8,452
A	2	22,745
B	2	147,775
AxB	4	21,126
Between Plots [Within AxB]	171	6,359
Within Plot	900	1,545
T	5	162,245
TxA	10	22,739
TxB	10	0,888
TxAxB	20	0,415
TxPlots [Within AxB]	855	0,391
Total	1079	2,691