MISSING DATA PROCEDURES: A COMPARATIVE STUDY (PART 2)

Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture
Washington, D.C. 20250

By Barry L. Ford
Sampling Studies Section
Sample Survey Research Branch
Statistical Research Division

MISSING DATA PROCEDURES:
A COMPARATIVE STUDY
(PART 2)

BARRY L. FORD
Sampling Studies Section
Statistical Research Division

Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture
Washington, D.C.
June, 1978

TABLE OF CONTENTS

SUMMARY AND CONCLUSIONS

The missing data procedure now used by ESCS for list frame surveys is to 1) delete the refusals and inaccessibles and reduce the sample size accordingly, and 2) have the statistician edit in values for single missing items. This procedure assumes that the missing data follow the same distribution as the reported data. To improve upon this assumption and to provide consistency in editing the single missing items, this study examines six missing data procedures. All six procedures rely upon control data of a high quality. Although this high quality is not important when the missing data are single missing items, it is extremely important to the missing data procedures when there are many refusals and inaccessibles. This study shows that better control data are needed before ESCS can replace the operational procedure.

Better control data implies larger correlations between the control variable and the variables reported on the questionnaire. The correlations within the data set used in this study are approximately 0.30 but evidence in this report indicates that correlations should be approximately 0.60 before there is a notable improvement over the operational missing data procedure. Artificial variables having large correlations with the control variable are incorporated into this report to compare procedures under the hypothesis that better control data can be obtained.

Of the six procedures studied in this experiment, three are slightly superior if the control data is adequate. These three candidates are the ratio procedure using balanced repeated replications (BRR), the hot deck procedure, and the hot deck procedure using BRR. (When the BRR technique is integrated into a missing data procedure, the product gives maximum insurance against potential biases due to replicate size while it simultaneously gives unbiased estimates of variance.) These procedures are recommended because of their simplicity or their statistical efficiency.

The statistical results in this report reveal no significant differences in the direct expansions from the missing data procedures <u>except</u> for the expected farrowing questions. The farrowing questions, however, present contradictory evidence. The hot deck procedure yields the <u>most</u> accurate estimates of the number of expected farrowings in the first quarter and the second <u>least</u> accurate estimates of the number of expected farrowing in the second quarter. Thus, the farrowing questions need further investigation using current data from several states to resolve the contradiction.

Because the statistical effects of the missing data procedures are only slightly different, non-statistical considerations should also be considered:

1. Initialization: the hot deck procedure requires initialization while the hot deck procedure (BRR) and the ratio procedure (BRR) do not.

2. Structure of the data set: the hot deck procedure requires a randomly ordered data set while the other two procedures require a specific, fixed order to the data set.

Both statistical and non-statistical comparisons will be crucial in a final decision. However, the first priority is the improvement of control data. The quality of the control data in most multiple frame states is unknown. The correlations between control data and reported data should be monitored in these states. Control data which is adequate for stratification may not be adequate for use with a missing data procedure. Research on obtaining and constructing better control numbers should also be planned. For example, several control variables may be more efficient than just one. Good control information is necessary for a good missing data procedure.

INTRODUCTION

A previous report, "Missing Data Procedures: A Comparative Study" (herein referred to as Part 1), focused on the problem of data missing from a list frame sample for a hog survey. The main purpose of this report was to compare three types of procedures which adjusted the estimates of the total number of hogs for missing records (i.e. refusals and inaccessibles). The three procedures--ratio, regression and hot deck--did not yield significant differences in the direct expansions of the total number of hogs. The lack of significance may have been due to the low correlations between the control information and the reported data. However, the hot deck procedure did yield more biased estimates of variance than the other procedures. This bias was the only statistical reason for concluding that the ratio and regression procedures were better than the hot deck procedure.

This report, Part 2, is still a comparison of the missing data procedures, but the comparison is made under conditions which are quite different from the conditions in Part 1. The changes are an enlargement of the capabilities of the missing data procedures, an investigation of a more accurate method of variance estimation for these procedures, and a simulation of the effect of better control data on these procedures.

THE PROCEDURES

A. The "Complete" Procedure

The "complete" procedure denotes the process of making estimates when the data set has no missing items. Although the "complete" procedure can not be used in an actual situation, it *can* often be used in simulation analysis. In this report its estimates may be regarded as the "true" sample values. Thus, one judgement of the quality of other missing data procedures is how close their estimates are to the estimates from the "complete" procedure.

B. The "Reported" Procedure

The "reported" procedure yields estimates by simply ignoring the missing data and reducing the sample size accordingly. Only the reported data set is expanded to make estimates. This procedure implies the assumption that the missing data are distributed the same as the reported data. The "reported" procedure is currently used by ESCS for refusals and inaccessibles in most list frame surveys. For a single missing item on a record the statistician imputes a number. Thus, the "reported" procedure is an approximation to the current, operational procedure.

Most of the probability surveys of ESCS are stratified. Stratification changes the above assumption to the assumption that the missing data follow the same distribution as the reported data *within each stratum*. Therefore, stratification gives a great deal of power to the "reported" procedures. If stratification is accurate and the number of strata is large, estimates from

this procedure will only be slightly affected by nonresponse bias. Accurate stratification implies a control variable which is highly correlated with the sample data. Thus, this procedure needs good control data to be effective. In fact, all of the following missing data procedures have this need.

C. The Ratio Procedure

The ratio procedure studied in this report is an adaption of the double sampling ratio estimator:

$$y_{ratio} = \frac{y^*}{x^*} x' + y_o \qquad (1.1)$$

where:

y^* is the total of a specific item, y, using only the records where both x and y were reported

x^* is the total of an auxiliary variable x, using only the records where both x and y were reported

x' is the total of the auxiliary variable, x, using all the records where x was reported

y_o is the total from the records in which only y was reported.

For missing records there is only one possible auxiliary variable--the control number. For example, one may wish to estimate the total number of hogs. Then y^* would be the total number of hogs from the reported data and x^* would be the total of the control variable for the reported data. Also, x' would be the total of the control variable for the entire sample, including complete and missing records. If only missing records are considered, y_o is zero.

The computer program to execute the ratio procedure has been enlarged from the previous study to include using information from partially complete records. The computer program takes a specific variable, y, and within each stratum finds the most highly correlated variable, x. The program then finds there are k records which have either the y or x variable reported. Suppose k = 10, "+" indicates reported and "-" indicates missing. One might have:

	y	x
1:	+	+
2:	+	+
3:	+	+
4:	+	+
5:	+	+
6:	+	+
7:	-	+
8:	-	+
9:	+	-
10:	+	-

The program would take y*, the total of the first six y values (both variables reported) and multiply it by x', the total of the eight reported x values and divide by x*, the total of the first six x values (both variables reported). The result would then have y_o, the total of the last two y values (only y reported), added to it.

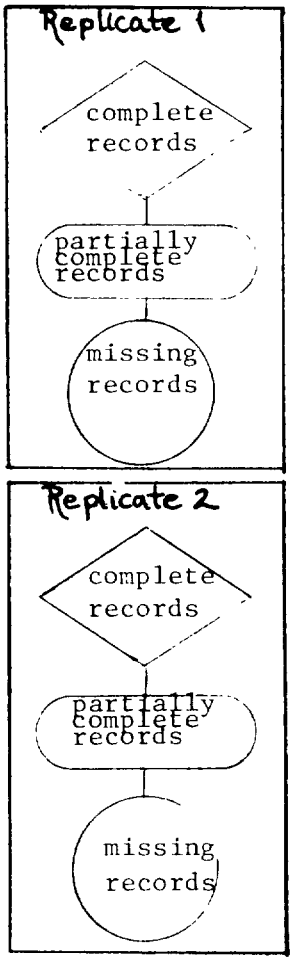
For example, if y = "farrowings in the last quarter", the program might find that x = "sows and gilts for breeding" is the most highly correlated variable. The program then looks at the records and finds there are six records with either x or y reported:

	y	x
1:	1	8
2:	4	12
3:	3	6
4:	4	10
5:	-	12
6:	4	-

The sum of the y values where both x and y are reported (y* = 12) is multiplied by the sum of the reported x values (x' = 48) and divided by the sum of the x values where both x and y are reported (x* = 36). To this result the program adds the sum of the y values where only y was reported (y = 4). Therefore, the estimate of the total number of "farrowings in the last quarter" from the complete and partial records is $y_{ratio} = 20$. This number is then adjusted by the control data to account for the missing records.

Once the most highly correlated variable (x) is found, there must exist records where x is reported when y is not. (If no such records are present one has $x' = x^*$ in equation (1.1), and thus, y_{ratio} simply equals the total of all reported y values.) When no such records exists, the second most highly correlated variable is found. If the same problem exists, then the control variable is used as the x variable.

To use the ratio procedure the data must enter the computer program in a specific order. Within each replicate the complete records must come first, the partially complete records second and the missing records last. For example, if there are two replicates, the data set is structured within each stratum as shown on the next page.



Within each of the three types of records, the records can be in any order. This data structure allows the computer program to make estimates with only one pass of the data.

D. Regression Procedure

The regression procedure studied in this report is an adaptation of the double sampling regression estimator:

$$y_{regr} = y^* + \hat{\beta} (x' - x^*) + y_o \qquad (1.2)$$

where x', x^*, y^* and y_o are defined as they were in the ratio procedure, and $\hat{\beta}$ is an estimator of the linear regression coefficient between x and y. Actually the ratio and regression estimators are two members of the class of linear estimators. Thus, they are closely related. In turn the computer programs executing the ratio and regression procedures are very similar except that one program uses equation (1.1) and the other uses (1.2). Therefore, a detailed explanation of the regression procedure is not given.

E. Hot Deck Procedure

A missing data procedure common to many large scale surveys is the hot deck procedure. For example, it is used by the Bureau of the Census and by Statistics Canada. Indeed, ESCS has used a type of hot deck procedure on data from labor surveys. Regardless of the particular survey, the hot deck procedure is basically a substitution process. Information from the reported data is substituted for the missing records. However, there are many types of hot deck procedures--from the simple to the sophisticated.

The hot deck procedure analyzed in this report is performed by a computer program designed by Norman Beller and written by Hugh Bynum in 1971 for use on the hog multiple frame surveys in Nebraska. The hot deck procedure goes through the following steps:

1: The records are randomly ordered within the strata, and then one record is selected at a time from this random order.

2: If there are reported values on the record, these reported values are entered into storage locations determined by the value of the control number. These storage locations are based on much smaller breakdowns of the control number than the breakdown used for stratification. (The hot deck program has the capability of using a geographic variable based on the crop reporting district although that capability is not used in this study.)

3: Each storage location contains a moving average--when a new value is entered, it is added to two times the old value and the resulting sum is divided by 3, i.e., $\left(\frac{2x_{old} + x_{new}}{3}\right)$.

4: If a record has a missing value, the storage location appropriate to the control number of that record is selected , and the value in the strorage location is substituted for the missing value.

The hot deck procedure requires initial values in the storage locations in case the first record has missing values. These initial values can be obtained from a previous multiple frame survey.

One should note that the major difference among hot deck programs is caused by step 4 above. The value kept in a storage location need not be a moving average but may be of almost any form--from a random, reported value to the mean of the reported values.

This hot deck program exploits relationships in the data in order to substitute for two farrowing items--first quarter intentions and second quarter intentions. These two items are handled in the same way so this report will explain the procedure for only the first quarter intentions item. The hot deck program builds a two-dimensional table of storage locations. One dimension of the table is the total number of hogs and the other dimension is the number sows farrowed this quarter. Every complete record is classified

into this table, and the percentage of the sows and gilts for breeding which will farrow next quarter is entered into the moving average in the storage location.

Now suppose a record has the sows farrowed this quarter and the total number of hogs reported on the questionnaire but lacks the first quarter intentions value. The hot deck takes the two reported values, finds the appropriate storage location in the two-dimensional table and uses the percentage in the storage location to calculate a first quarter intentions value for the missing item.

Example: A record has total number of hogs = 100, sows farrowed in the current quarter = 40, sows and gilts for breeding = 80 and the first quarter intentions is missing. The hot deck program goes to the storage location marked with an "x" in the two-dimensional array below and retrieves the percentage stored there. Suppose that percentage is 0.45. Then 0.45 x 80 = 36 is used as the first quarter intentions on the record.

		Total Number of Hogs				
		0-30	31-60	61-100	100-200	over 200
	0-25					
Sows	26-50			X		
Farrowed	51-100					
This	101-200					
Quarter	over 200					

Similarly, if the record is complete, the percentage--first quarter intentions : sows and gilts for breeding--is entered into the moving average in the "x" storage location.

The computer program executing the hot deck procedure only imputes for three items on a partially complete record--the first quarter intentions, the second quarter intentions, and the total number of hogs. Thus, only these three items are used in comparing the hot deck procedure with the other missing data procedures.

The data set must enter into the computer program in a random order within each stratum. This structure is caused by the fact that the order of the records affects the values imputed. For instance, if the missing records are grouped together, they tend to get similar imputed values. The similar values cause an artificial decrease in the estimation of standard errors, i.e. a downward bias. If a random ordering is used, this bias decreases.

REPLICATION

Although the primary emphasis of Part 2 is the accurate estimates of means or totals, an important (though secondary) goal is the accurate estimation of standard errors. The previous report, Part 1, listed many reasons why the missing data procedures may not yield accurate estimates of standard errors. With the extenstion of these procedures to partially missing records this problem is even more serious. The computer programs have allowed the missing data procedures to become so complex that formulas for the standard errors are extremely difficult and perhaps impossible to derive. However, a sampling strategy which allows the computation of estimates of standard errors is replication.

Replication is a strategy in which the sample is randomly divided into r subsamples (called replicates)--each subsample having the ability to produce an estimate of the population total. By using the variablility between these estimates from the different replicates one can estimate the standard error. If $T_1, T_2, \ldots T_r$ are the estimates of the population total from the r independent replicates, then the final population estimate is simply the average of the replicate estimates, i.e.:

$$T = \frac{\sum_{i=1}^{r} T_i}{r} = \frac{T_1 + T_2 + \ldots + T_r}{r} \qquad (1.3)$$

The estimated standard error of T, S(T) is:

$$S(T) = \left[\frac{\sum_{i=1}^{r} (T_i - T)^2}{r(r-1)}\right]^{\frac{1}{2}} \qquad (1.4)$$

Example: The stratification has a simple random sample of size 21. The statistician decides to form 3 replicates of 7 units each. His sample design might then be:

		Replicate		
		1	2	3
Unit	1	6	8	10
	2	4	10	14
	3	6	4	6
	4	10	6	10
	5	6	4	4
	6	8	2	8
	7	2	1	18

where the values inside the box (6, 4, 6, 10 etc.) are observed values on the sample unit.

To calculate an estimate of the population total and its standard error one must calculate the mean of each replicate and multiply by an expansion factor. Suppose the expansion factor is to:

Replicate

1	2	3
$T_1 = \frac{42}{7}(10) = 60$	$T_2 = \frac{35}{7}(10) = 50$	$T_3 = \frac{70}{7}(10) = 100$

The final estimates of the population total is the average of the replicate estiamtes:

$$T = \frac{\sum_{i=1}^{3} T_i}{3} = \frac{60 + 50 + 100}{3} = 70.$$

The standard error of T is:

$$S(T) = \left[\frac{\sum_{i=1}^{3} (T_i - T)^2}{3(3-1)}\right]^{\frac{1}{2}} = \left[\frac{(60-70)^2 + (50+70)^2 + (100-70)^2}{6}\right]^{\frac{1}{2}}$$

$$= \left[\frac{100 + 400 + 900}{6}\right]^{\frac{1}{2}} = 15.3$$

The most difficult problem when replicating a sample is to decide how many replicates are needed. If the total sample size is 100, should there be 10 replicates of size 10, 20 replicates of size 5, or 50 replicates of size 2?

When replicating a sample in which one has fixed the total sample size, there are two major forces at work. One of these is the decrease in the stability of the standard error estimates when the number of replicates decreases. In other words, if one uses a small number of replicates, there will be wide fluctuations in the standard error estimates over many surveys. Opposing this force is a decrease in the size of many biases of estimated means and totals when the number of replicates decreases. Thus, in the above example 2 replicates of size 50 probably yields an unstable estimate of the standard error but a small bias in its estimate of the total.

It is well known (2) that the ratio and regression estimators have biases which are affected by a small replicate size. Replicates of a small size lead to severe biases in these two procedures. Thus, these procedures depend on a large replicate size to produce accurate estimates of means or totals. The hot deck procedure also depends on a large replicate size in order to be effective. If the replicate size is small, cells in the two-way tables must be collapsed. Thus, the efficiency of the hot deck is impaired.

One of the newest techniques of replication in a stratified design is called balanced repeated replications (BRR). The BRR technique (5, 6) allows the replicate size to remain large (i.e., the number of replicates to remain small) while keeping a fairly stable estimate of the standard error. In fact, the BRR technique requires only two replications per stratum. Using orthogonal vectors, the BRR technique uses the two replicates per stratum to compute population estimates and standard errors.

Example: Suppose there are 3 strata containing the following sample values which are assigned to two replicates:

		Stratum 1	Stratum 2		Stratum 3			
Replicate	1	6	4	6	10	6	8	6
	2	8	10	4	6	4	7	1

The average of each replicate in each stratum is:

		Stratum 1	Stratum 2	Stratum 3
Replicate	1	6	5	7.5
	2	8	7	4.5

One must now obtain 3 orthogonal vectors whose length will be the number of estimates of the population total. For instance, one finds (7) that 3 orthogonal vectors of length 8 are:

		Stratum 1	Stratum 2	Stratum 3
Estimate	1	+	+	+
	2	-	+	+
	3	-	-	+
	4	+	-	-
	5	-	+	-
	6	+	-	+
	7	+	+	-
	8	-	-	-

A "+" signifies the use of replicate 1 and a "-" signifies the use of replicate 2. Thus, from this diagram the fourth estimate of the population total, t_4, is made by using the first replicate of stratum 1, the second replicate of stratum 2 and the second replicate of stratum 3. All eight estimates of the population total are:

	1	+	+	+	$t_1 = 6N_1 + 5N_2 + 7.5N_3$
	2	-	+	+	$t_2 = 8N_1 + 5N_2 + 7.5N_3$
	3	-	-	+	$t_3 = 8N_1 + 7N_2 + 7.5N_3$
Estimate	4	+	-	-	$t_4 = 6N_1 + 7N_2 + 4.5N_3$
	5	-	+	-	$t_5 = 8N_1 + 5N_2 + 4.5N_3$
	6	+	-	+	$t_6 = 6N_1 + 7N_2 + 7.5N_3$
	7	+	+	-	$t_7 = 6N_1 + 5N_2 + 4.5N_3$
	8	-	-	-	$t_8 = 8N_1 + 7N_2 + 4.5N_3$

where N_1, N_2 and N_3 are population sizes in strata 1, 2 and 3 respectively.

Final population estimates are:

$$T = \frac{\sum_{i=1}^{8} t_i}{8} = \frac{t_1 + t_2 + \ldots + t_8}{8}$$

$$S(T) = \left[\frac{\sum_{i=1}^{8} (t_i - T)^2}{8}\right]^{\frac{1}{2}} = \left[\frac{(t_1 - T)^2 + (t_2 - T)^2 + \ldots + (t_8 - T)^2}{8}\right]^{\frac{1}{2}}$$

If $N_1 = N_2 = N_3 = 100$:

$t_1 = 1850$

$t_2 = 2050$

$t_3 = 2250$

$t_4 = 1750$

$t_5 = 1750$

$t_6 = 2050$

$t_7 = 1550$

$t_8 = 1950$

$T = 1900$

$S(T) = 206$

When there are only 2 replicates per stratum (each replicate containing half of the total sample), the estimates of means or totals may still be less accurate than the estimates calculated when no replication is used. For this reason one may calculate an estimate of a mean or total without using replication and then estimate the standard error by using the BRR technique. This approach is often adopted (4), but in the circumstances of the ESCS problem it requires an extra pass of the data. Thus, although the costs may be prohibitive, it is a option to remember.

ANALYSIS

A. Purpose

The purpose of this analysis is to find the best missing data procedure. The best missing data procedure:

1: yields direct expansions which are the closest to the direct expansions when the data is complete.

2: yields direct expansions which are mathematically unbiased.

3: yields accurate and small standard error estimates.

B. Data

The data in this report are the complete records from the list frame hog survey of one state. There are 1081 records. Originally there were nine strata, but two strata are ignored in this study because their control data are zeroes. As noted in Part 1, a control number of zero is useless to any of these missing data procedures.

This simulation experiment involved randomly choosing 20 percent of the records to be missing. Thus, all of the data items on these records were deleted except for the control data. Another 15-20 percent of the records were chosen to be partially incomplete. One of the data items--total number of hogs, expected farrowings in the first quarter, expected farrowings in the second quarter, previous farrowings, pigs still on hand, or pigs sold last quarter--was randomly deleted from each of these records.

The records designated as missing or partially incomplete were chosen randomly but had unequal probabilities of selection which were proportional to the total number of hogs. The result of this process was that the records of the larger hog operations were more likely to be selected as incomplete or missing than those of the smaller hog operations. Five data sets were created from the complete data set. Each of these five data sets had different records chosen to be incomplete or missing.

Control data play an important role in the efficiency of a missing data procedure. This importance is particularly true for missing records because the control variable is the only information available. The control data in the test state is poor, i.e. the correlations between the control variable and the total number of hogs is about 0.30. Figure 1 graphically illustrates this fact. The stratification does not really separate the records into homogenous groups. It does seem to separate the records into groups which have greater dispersion as the number representing the stratum increases. For example, stratum 5 has more dispersion than stratum 2. Thus, the points on the graph have a triangular shape.

The most striking aspect of Figure 1 is that within each stratum there is little relationship between the control variable and the total number of hogs. One notices, for example, that in stratum 5 the points are randomly scattered across the page. This quality of the control data is not so bad if one simply wants the control variable to separate the population into four or five broad groups. However, when one uses the control variable in missing data procedures in order to adjust for refusals and inaccessibles, the quality needs to be a great deal better.

Figure 2 displays the relationship between the control variable and the total number of hogs when the correlation between these two variables is approximately 0.90. Not only is there a strong linear relationship within each stratum, but there is also a strong separation of the data into one homogenous group within each stratum. It does not matter so much that these groups overlap somewhat with regard to the total number of hogs. What does matter in Figure 2 is that the data in each stratum is a compact unit with a linear trend.

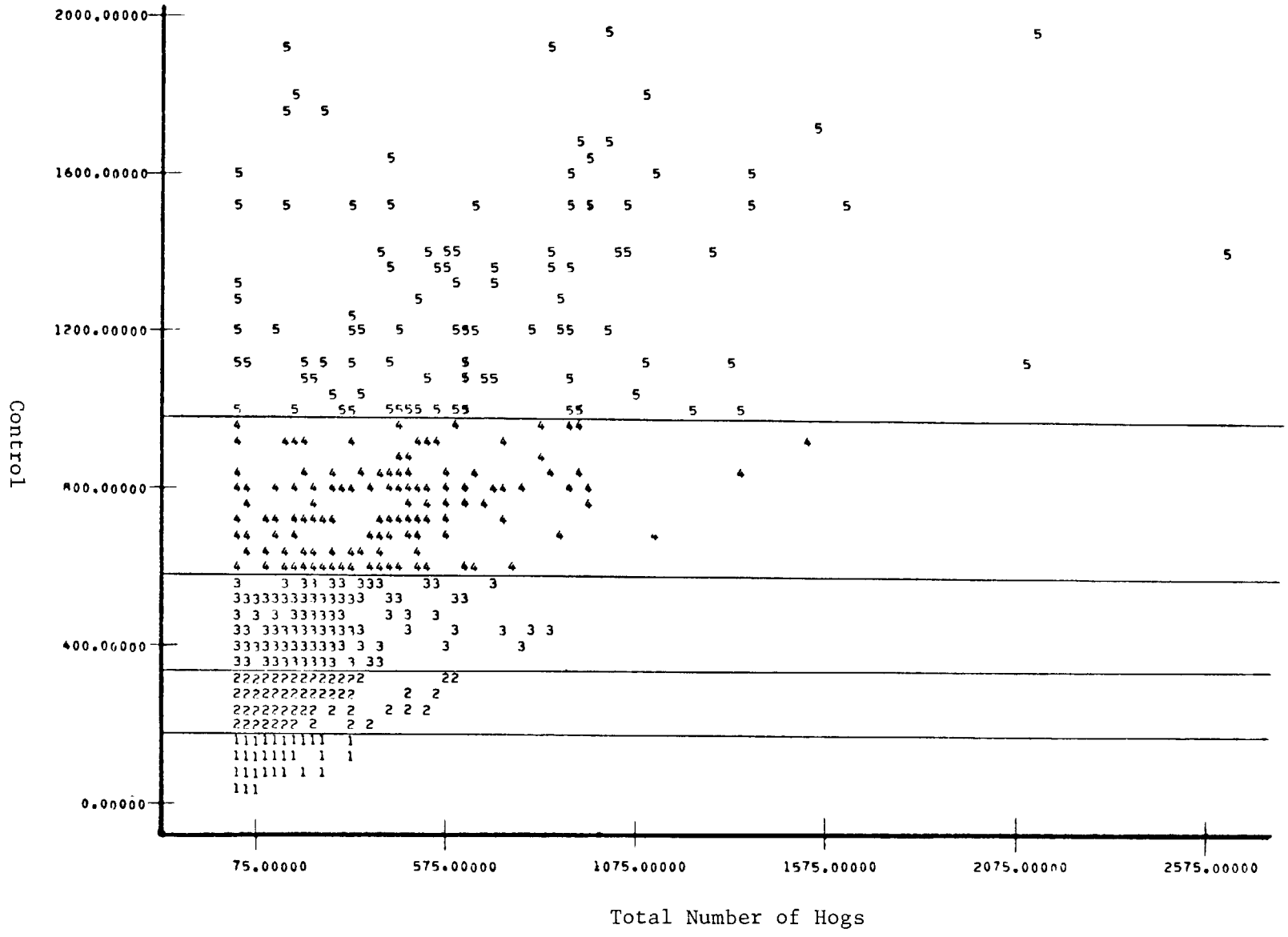


Figure 1: A graph of the total number of hogs and the value of the control variable in five strata. Within each stratum correlations between these two variables are approximately 0.30. Each point on the graph is represented by a number from one to five that identifies the stratum to which that point belongs.

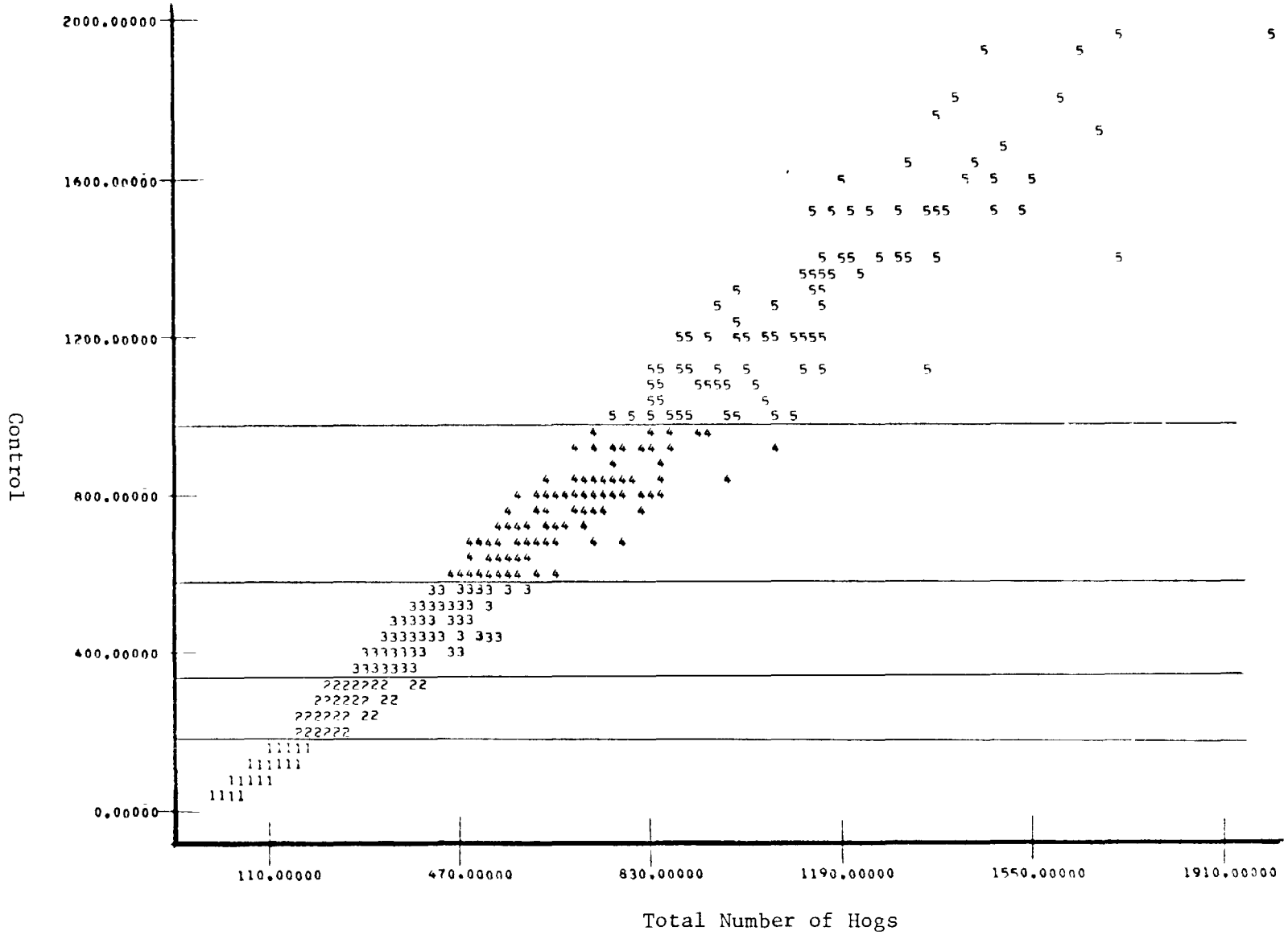


Figure 2: A graph of the control variable and an aritifical variable. The artificial variable was calculated so that correlations with the control variable are approximately 0.90 within each stratum. Each point on the graph is represented by a number from one to five that identifies the stratum to which that point belongs.

C. Results

Table 1 displays the direct expansion from the missing data procedures. Table 2 displays the results in percentage terms, i.e. the direct expansions from each missing data procedure are divided by the direct expansion of the "reported" procedure. For example, if the data set is complete, the direct expansion of the total number hogs is 1.13 times the direct expansion from the "reported" procedure. Therefore, for the total number of hogs a ratio of 1.00 indicates no improvement over the "reported" procedure while 1.13 indicates perfect agreement with the "truth"--the direct expansion from the complete data set. One can note from column 1 in the table that there is little improvement in the total number of hogs when using any of the missing data procedures. An analyis of variance found no significant differences in the estimates of the total number of hogs.

A small correlation (a weighted average of 0.30) between the total number hogs and the control variable is primarily responsible for the poor improvements of the missing data procedures. To change the values of the control variable in this experiment so that the correlation is larger involves substantial changes in the computer programs. Therefore, two new variables--pseud 1 and pseud 2--were created to have larger correlations with the control variable. When the data set is complete, pseud 1 and pseud 2 yield the same direct expansions as the number of total hogs, but they provide correlations of 0.64 and 0.87 respectively. Thus, the effects of larger correlations on the missing data procedure can be studied.

The direct expansions of pseud 1 shown in Table 2 are 1.03 or 1.04 times larger than the direct expansion of the "reported" procedure. Duncan's multiple comparison test reveals a significant improvement over the "reported" procedure by all the missing data procedures. Thus, when the correlation between the control number and a variable is 0.64, the "reported" procedure should definitely be discarded in favor of another missing data procedure. The next question is which procedure does one choose. Duncan's multiple comparison test reveals no significant differences among the estimates of pseud 1 from the other procedures. The question can not be answered at this point.

Pseud 2 increases the average correlation to 0.87. Direct expansions of pseud 2 from the procedures are 1.06 or 1.07 times larger than the direct expansion from the "reported" procedure. Duncan's multiple comparison test reveals the same result for pseud 2 as for pseud 1. All of the other missing data procedures yield significantly better estimates of pseud 2 than the "reported" procedure. However, there are no significant differences among the estimates from these other procedures.

Table 1: Direct expansions calculated from the missing data procedures.

PROCEDURES	VARIABLES				
	Total Hogs (1,000,000)	Pseud 1** (1,000,000)	Pseud 2** (1,000,000)	Expected Farrowings : First Quarter (100,000)	Expected Farrowings : Second Quarter (100,000)
"Complete"	10.584	10.584	10.584	5.011	6.548
"Reported"	9.364	9.364	9.364	4.457	5.820
Hot Deck	9.418	9.750	10.058	4.850	5.910
Hot Deck : BRR*	9.389	9.646	10.008	4.823	6.017
Ratio	9.498	9.661	10.040	4.695	6.115
Ratio : BRR	9.412	9.639	9.958	4.613	5.983
Regression	9.483	9.633	9.957	4.712	6.064
Regression : BRR	9.393	9.625	9.958	4.610	5.981

*BRR refers to the technique of balanced repeated replications.

**The correlation between the total number of hogs and the control variable is 0.30. Pseud 1 and Pseud 2 are artificial variables created to test the effects of the missing data procedures when the correlations with the control variable are 0.64 and 0.87 respectively.

Table 2: Direct expansions of the missing data procedures divided by the direct expansion calculated from the reported procedure.

PROCEDURES	VARIABLES				
	Total Hogs	Pseud 1**	Pseud 2**	Expected Farrowings : First Quarter	Expected Farrowings : Second Quarter
"Complete	1.13	1.13	1.13	1.12	1.13
"Reported"	1.00	1.00	1.00	1.00	1.00
Hot Deck	1.01	1.04	1.07	1.09	1.02
Hot Deck : BRR*	1.00	1.03	1.07	1.08	1.03
Ratio	1.01	1.03	1.07	1.05	1.05
Ratio : BRR	1.01	1.03	1.06	1.04	1.03
Regression	1.01	1.03	1.06	1.06	1.04
Regression : BRR	1.00	1.03	1.06	1.03	1.03

* BRR refers to the technique of balanced repeated replications.

** The correlation between the total number of hogs and the control variable is 0.30. Pseud 1 and Pseud 2 are artificial variables created to test the effects of the missing data procedures when the correlations with the control variable are 0.64 and 0.87 respectively.

The number of expected farrowings during the first quarter and the number during the second quarter provide contradictory information. Only the hot deck procedure yields a direct expansion of the expected farrowings during the first quarter which is a significant improvement over the "reported" procedure. (One should note that the hot deck procedure (BRR) is almost significant at the 5% level.) Indeed, the 1.09 ratio is so remarkable that the tabular method used by the hot deck computer program to impute for expected farrowings (see discussion of the hot deck procedure) appears to be the best approach to the missing data problem. However, the number of expected farrowings during the second quarter gives a conflicting result. The direct expansion using the hot deck procedure is not significantly different from the direct expansion of the reported procedure. Such contradiction may be due to the specific data set, to the procedure, to hog data in general, or many other reasons. Applying the missing data procedures to current data from several states should help to clarify this contradiction.

One might wonder why there is more improvement in the expected farrowing questions than in the total number of hogs. The question is even more vexing when one realizes that the correlation between the number of expected farrowings and the control number is no larger than the correlation between the total number of hogs and the control number. The explanation is complicated. The total number of hogs on a record is composed of several weight and breeding subclasses. The computer program for the hot deck procedure only imputes for the total number and not for any of the subclasses. Thus, if one has any of the subclass information when the total is missing, it can not be used in calculating a value for the total. This kink is only a result of the computer program and not true of the hot deck procedure in general. It was too time consuming for this project to write a new computer program for a procedure as complex as the hot deck procedure. The ratio and regression procedures can use the subclass information but were not allowed this benefit in order to have a fairer comparison with the hot deck procedure. The expected farrowing questions on all the procedures can and did use the subclass information. The ability is important because the number of sows for breeding is closely related to the number of expected farrowings. The result is more dramatic improvements in expected farrowings than in the total number of hogs.

One must remember that the correlation between a variable and the control number is important because the only bit of information available for missing records is the control variable. The total number of hogs, pseud 1 and pseud 2 illustrate this point in Table 2, and Figure 3 illustrates the fact graphically. Given the percentage of missing items in this study, one could chart the improvement in the direct expansions of a variable as a result of the increased correlation with the control number. Figure 4 displays this relationship for the non-operational procedures in general. As one can see from the graph, correlations should be at least are around 0.60 before improvements can be worthwhile.

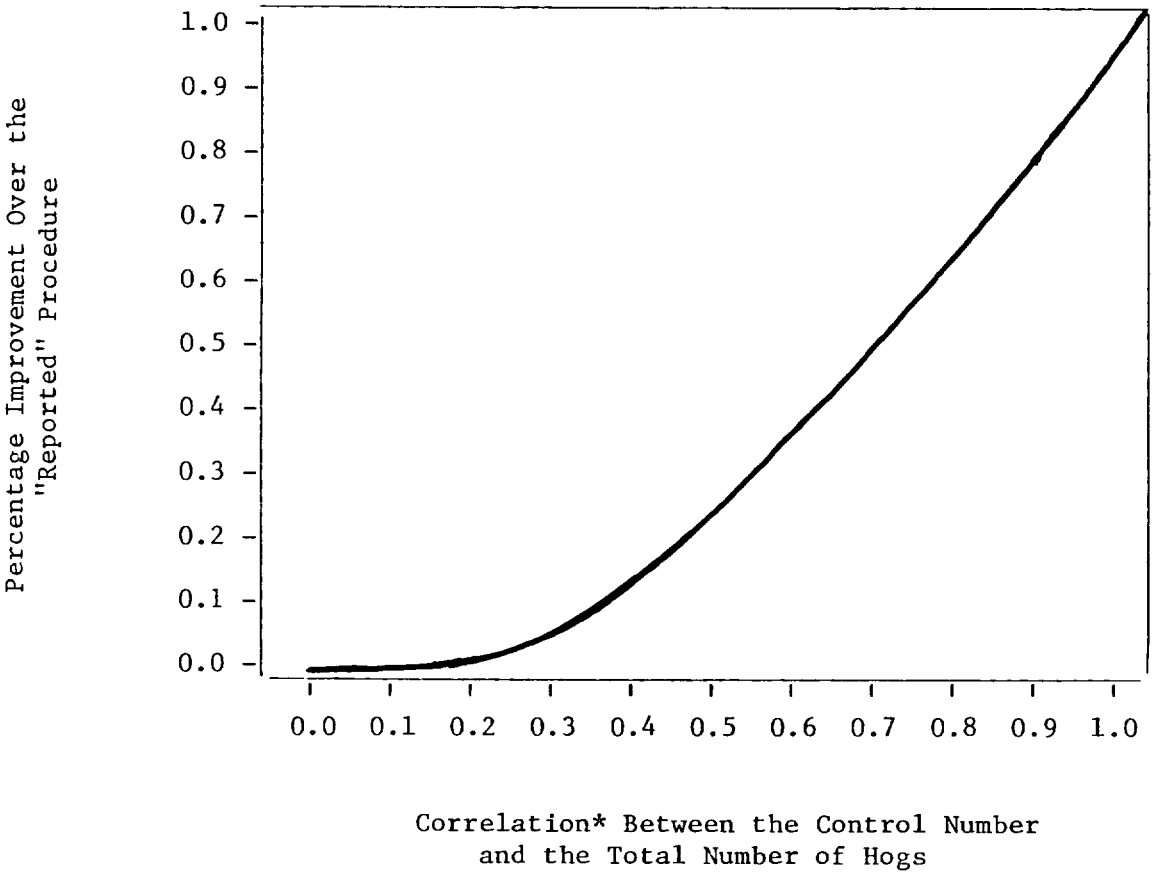


Figure 4: A graph displaying the percentage improvement in the direct expansion of the total number of hogs versus the correlation between the control number and the total number of hogs. The percentage improvement is in the direct expansion from a missing data procedure over the direct expansion from the "reported" procedure. Thus "0" indicates no improvement over the "reported procedure" while "1" indicates the most possible improvement.

*The correlation in this graph refers to a weighted correlation--an overall indicator of correlation. The weighted correlation equals $\sum_{i=1}^{L} w_i \rho_i$, where L = the number of strata, w_i = the weight for stratum i and ρ_i = the correlation in stratum i.

The discussion of the computer program for the hot deck procedure noted a requirement of initial values for the hot deck procedure. These initial values have an effect on the direct expansion produced. Figure 5 displays the difference in improvement when the reported data set is used to initialize it. Of course, using the complete data set is not a viable alternative from which to choose. This graph merely illustrates the fact that the improvement in the estimates from a hot deck procedure can change drastically according to the quality of the initial estimates. Initialization may be based on the data from previous surveys and under those circumstances tends to pull the current estimates towards the previous estimates.

The technique of balanced repeated replications (BRR) was applied to the hot deck, ratio and regression procedures for different reasons. In the case of the hot deck procedure it was hoped that the BRR technique would yield unbiased estimates of the standard error without affecting the direct expansions. Table 2 shows that the BRR technique hardly affects the direct expansions of the hot deck procedure. Estimates of the standard errors are shown in Table 3; their ratios to the standard error from the "reported" procedure are shown in Table 4. One can see in Table 4 that the hot deck procedure seriously understates the estimates of the standard error. The variability in the standard error estimates prevents the F-test of analysis of variance from identifying the standard error estimates as significantly different. However, the experiment in Part 1 had a sample size which enabled one to draw the conclusion that the hot deck procedure yields esitmates of standard error which are consistently less than the estimated standard error of the complete data set. Thus, the standard errors of the hot deck procedure must have a downward bias. Using the BRR technique, one alleviates this bias.

By using the BRR technique on the hot deck procedure one can also eliminate the need for initialization. (This advantage was not used in the hot deck (BRR) procedure in Part 2.) The direct expansions will then behave as though they were initialized with the reported data as in Figure 5. Furthermore, the BRR technqiue also makes it possible to eliminate the need for randomly ordering the data set. The hot deck (BRR) procedure can have data entered in the same format as the ratio and regression procedures.

The BRR technique was also applied to the ratio and regression procedures. These two procedures already use independent replicates of a small size to produce estimates of the standard error. The small replicate size may cause an upward bias in the direct expansions. The BRR technique allows a large increase in the size of the replicates and thus, decreases this bias. Table 2 reveals little difference in the direct expansions although there does seem to be a small upward bias in the direct expansions of the ratio and regression procedures. However, this bias is so small that a multivariate test could locate no difference between the ratio procedure and the ratio (BRR) procedure or between the regression procedure and the regression (BRR) procedure.

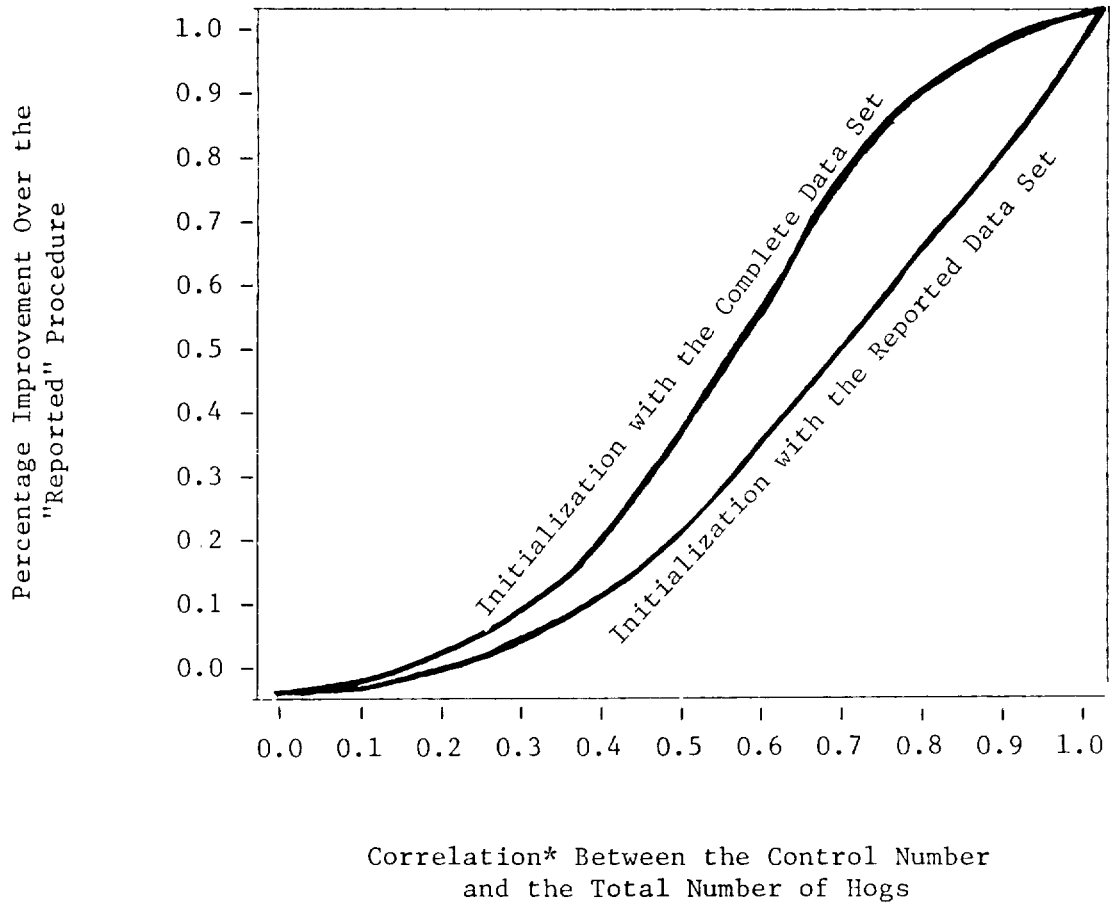


Figure 5: A graph displaying the effect of initial values on the direct expansions from the hot deck procedure. One method of intitialization is the reported data set and one method is the complete data set. The effects of these two methods are charted with respect to the correlation between the control number and the total number of hogs and with respect to the percentage improvement in the direct expansion from the hot deck procedure over the "reported" procedure.

* The correlation in this graph refers to a weighted correlation--an overall indicator of correlation. The weighted correlation equals $\sum_{i=1}^{L} w_i \rho_i$, where L = the number of strata, w_i = the weight for stratum i and ρ_i = the correlation in stratum i.

Table 3: Estimates of standard errors calculated from different missing data procedures.

PROCEDURES	VARIABLES		
	Total Hogs (100,000)	Expected Farrowings : First Quarter (10,000)	Expected Farrowings : Second Quarter (10,000)
"Complete"	2.982	2.596	2.817
"Reported"	3.076	2.805	2.934
Hot Deck	2.496	2.478	2.451
Hot Deck : BRR*	3.018	2.841	3.635
Ratio	3.216	3.409	3.586
Ratio : BRR	2.962	2.226	3.005
Regression	3.216	3.422	3.397
Regression : BRR	2.920	2.195	3.010

*BRR refers to the technique of balanced repeated replications.

Table 4: Standard error estimates divided by the standard error estimate calculated from the "reported" procedure.

PROCEDURES	VARIABLES		
	Total Hogs	Expected Farrowings : First Quarter	Expected Farrowings : Second Quarter
"Complete"	0.97	0.93	0.96
"Reported"	1.00	1.00	1.00
Hot Deck	0.81	0.88	0.84
Hot Deck : BRR*	0.98	1.01	1.24
Ratio	1.05	1.22	1.22
Ratio : BRR	0.96	0.80	1.02
Regression	1.05	1.22	1.16
Regression : BRR	0.95	0.78	1.03

*BRR refers to the technique of balanced repeated replications.

Table 4 shows that the BRR technique does improve the level of standard errors. Multivariate tests show that the esitmates of standard error for the hot deck, ratio and regression procedures are not significantly different at a 10% level from their BRR counterparts, but the significance levels are small. Significance levels are 0.18, 0.23 and 0.25.

The BRR method is an unbiased technique and a safer technique than the many independent replicates of the ratio and regression procedures. The BRR technique is also easy to use--the subsample in each stratum only needs to be divided in half. Thus, based on the evidence in Part 2 the ratio (BRR) and regression (BRR) procedures are considered better than the ratio and regression procedures.

Multivariate tests revealed no difference in the vector of estimates from the ratio (BRR) and the vector from regression (BRR) procedures. Since the ratio (BRR) procedure is slightly easier to compute, it is slightly better than the regression (BRR) procedure.

Applying the BRR technique to any procedure may result in instability of the standard error estimates (dicussed in the "Replication" selection). Table 4 demonstrates this fact through inconsistencies in the standard error estimates when the BRR technique is used. For example, the ratio of 1.24 for the hot deck procedure (BRR) appears as abnormally large, and 0.80 for the ratio procedure (BRR) appears as abnormally small. For the hot deck, ratio and regression procedures this instability is the only drawback to the BRR technique and probably not a serious one.

BIBLIOGRAPHY

1. Beyer, William H. (editor). CRC Handbook of Tables for Probability and Statistics. Chemical Rubber Company, 1968.

2. Cochran, William G. Sampling Techniques. John Wiley and Sons. 1968.

3. Ford, Barry L. Missing Data Procedures: A Comparative Study. 1976.

4. Frankel, Martin R. and Frankel, Lester R. "Some Recent Developments in Sample Survey Design," Journal of Marketing Research. Volume XIV. 1977.

5. Kish, Leslie and Frankel, Martin R. "Balanced Repeated Replications for Standard Errors," Journal of the American Statistical Association. Volume 65. 1970.

6. McCarthy, P.J. "Pseudo-replication: Half Samples" Journal of the International Statistical Institute. Volume 37. 1969.

7. Plackett, R.L. and Burnam, P.J. "The Design of Optimum Multifactorial Experiments," Biometrika. Volume 33. 1946.

APPENDIX

The experimental design used in this experiment is a randomized complete block design (i,e. a two-way analysis of variance with one observation per cell). The two factors in this design are A) a treatment effect due to the effect of a missing data procedure, and B) a block effect due to the random deletion of items and records from the sample. The treatment effect is fixed and the block effect is random. One observation per cell necessitates the assumption of no interaction between the two effects.

The seven levels of the treatment effect are:

1. the "reported" procedure
2. the ratio procedure
3. the ratio procedure using balanced repeated replications (BRR)
4. the regression procedure
5. the regression procedure using BRR
6. the hot deck procedure
7. the hot deck procedure using BRR

The estimates of the "complete" procedure, of course, do not vary no matter which items or records are deleted. Therefore, although the "complete" procedure is not a treatment effect, its estimates are used in the analysis as benchmarks to measure the improvement in the estimates of the other missing data procedures.

In order to apply the missing data procedures, the complete data set had 20 percent of its records deleted entirely, and approximately another 20 percent of its records had item deletions. As noted in the main body of the report, if the total number of hogs was deleted, then all of the weight and breeding subclasses were deleted. Deletions were made randomly, but records with a larger total number of hogs had a larger probability of being deleted. This deletion process was applied to the complete data set five times. The result was five incomplete data sets. Therefore, there are five levels to the block effect--each level corresponding to one of the incomplete data sets. Each missing data procedure was applied to each of these five data sets.

There are five dependent variables for which there are estimates:

1. the total number of hogs
2. pseud 1--an artificial variable having an average correlation of 0.64 with the control number
3. pseud 2--an artificial variable having an average correlation of 0.87 with the control variable
4. the number of expected farrowings in the first quarter
5. the number of expected farrowings in the second quarter

There are also two types of estimates for each variable--the direct expansion and the estimated standard error of the direct expansion. The estimated standard errors for pseud 1 and pseud 2 are not included in the analysis.

The model for this experimental design is:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where:

$i = 1, 2, \ldots, 7$
$j = 1, 2, \ldots, 5$
α_i = the effect of the ith missing data procedure
β_j = the effect of the jth incomplete data set
ε_{ij} = the error in the model associated with y_{ij} ($\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$)
y_{ij} = the value of a dependent variable when missing data procedure i is applied to data set j

Table 5-12 exhibit the appropriate analysis of variance table for this model. There are eight F-tests--a test on the direct expansion of each of the five variables and a test on the estimated standard errors of three variables. Within the context of a specific variable and a specific type of estimate, one tested:

H_o: There is no difference in the effects of the missing data procedures.
H_a: There is a difference in the effects of the missing data procedure.

Table 5: Analysis of variance on the direct expansions of the total number of hogs.

Source	Degrees of Freedom	Sum of Squares (10^{11})
A	6	7.447
B	4	68.106
Error	24	6.793
Total	34	143.486

F = 0.44
Significance Level = 0.85

Table 6: Analysis of variance on the direct expansions of pseud 1

Source	Degrees of Freedom	Sum of Squares (10^{11})
A	6	4.259
B	4	1.983
Error	24	1.772
Total	34	8.014

F = 9.61
Significance Level ≤ 0.00

Table 7: Analysis of variance on the direct expansions of pseud 2

Source	Degrees of Freedom	Sum of Squares (10^{11})
A	6	13.054
B	4	0.468
Error	24	1.143
Total	34	14.665

F = 45.70
significance Level ≤ 0.00

Table 8: Analysis of variance on the direct expansions of the number of expected farrowings in the first quarter.

Source	Degrees of Freedom	Sum of Squares (10^9)
A	6	5.617
B	4	10.689
Error	24	17.131
Total	34	33.437

F = 1.31
Significance Level = 0.29

Table 9: Analysis variance on the direct expansions of the number of expected farrowings on the second quarter.

Source	Degrees of Freedom	Sum of Squares (10^9)
A	6	2.797
B	4	2.590
Error	24	3.952
Total	34	9.339

F = 2.83
Significance Level = 0.03

Table 10: Analysis of variance on the estimated standard errors of the total number of hogs.

Source	Degrees of Freedom	Sum of Squares (10^{10})
A	6	1.798
B	4	5.553
Error	24	13.247
Total	34	20.598

F = 0.54
Significance Level = 0.77

Table 11: Analysis of variance on the estimated standard error of the number of expected farrowings in the first quarter.

Source	Degrees of Freedom	Sum of Squares (10^8)
A	6	7.755
B	4	5.290
Error	24	15.566
Total	34	28.611

F = 1.99
Significance Level = 0.11

Table 12: Analysis of variance on the estimated standard error of the number of expected farrowings in the second quarter.

Source	Degrees of Freedom	Sum of Squares (10^8)
A	6	5.315
B	4	2.936
Error	24	9.913
Total	34	18.164

F = 2.14
Significance Level = 0.09

The test statistic:

$$F = \frac{\text{Mean square of A}}{\text{Error mean square}}$$

follows an *F* - distribution with 6 and 24 degrees of freedom. The values of the F statistics and their significance levels are also in Tables 5-12.

The F-tests are not as important as the multiple comparison tests. This study uses Duncan's multiple comparison test to make pairwise tests of significance on estimates from the missing data procedures. In other words one uses Duncan's multiple comparison test to locate exactly which procedures yielded different direct expansions from the others. For a specific variable, *a*, two missing data procedures are significantly different if:

$$|\alpha_i - \alpha_j| \geq P \sqrt{\frac{2\ (\text{Error Mean Square})}{n}} \qquad (1.5)$$

where:

$_i$ = the average estimate of from the missing data procedure i

$_j$ = the average estimate from the missing data procedure j

P = the critical value for Duncan's test; found in Duncan's table (1, page 368)

n = the number of data sets = 5

(In this experiment Duncan's tests were carried out at a five percent significance level.)

Figures 4-11 show the results of Duncan's test. Each vertical line connects a group of procedures. Duncan's test says that the procedures within each group do not yield significantly different estimates. In other words, *two missing data procedures yield signficantly different estimates if they are not connected by a vertical line.* Although the "complete" procedure is not included in the test, its estimates are given in the figures for comparison.

Figure 6: The results of Duncan's multiple comparison test on the direct expansions of the number of total hogs.

Missing Data Procedures	Estimate : Estimate from "Reported" Procedure	Groupings
"Reported"	1.00	I
Hot Deck (BRR)	1.00	I
Regression (BRR)	1.00	I
Ratio (BRR)	1.01	I
Hot Deck	1.01	I
Regression	1.01	I
Patio	1.01	I
"Complete"	1.13	

Figure 7: The results of Duncan's multiple comparison test on the direct expansions of pseud 1.

Missing Data Procedure	Estimate : Estimate from "Reported" Procedure	Groupings
"Reported"	1.00	I
Regression (BRR)	1.03	I
Regression	1.03	I
Ratio (BRR)	1.03	I
Hot Deck	1.03	I
Ratio	1.03	I
Hot Deck	1.04	I
"Complete"	1.13	

Figure 8: The results of Duncan's multiple comparison test on the direct expansions of pseud 2.

Missing Data Procedure	Estimate ÷ Estimate from "Reported" Procedure	Groupings
"Reported"	1.00	I
Regression	1.06	I
Ratio (BRR)	1.06	I
Regression (BRR)	1.06	I
Hot Deck (BRR)	1.07	I
Ratio	1.07	I
Hot Deck	1.07	I
"Complete"	1.13	

Figure 9: The results of Duncan's multiple comparison test on the direct expansions of the number of expected farrowings in the first quarter.

Missing Data Procedure	Estimate ÷ Estimate from "Reported" Procedure	Groupings
"Reported"	1.00	I
Regression (BRR)	1.03	I
Ratio (BRR)	1.04	I
Ratio	1.05	I
Regression	1.06	I
Hot Deck (BRR)	1.08	I
Hot Deck	1.09	I
"Complete"	1.12	

Figure 10: The results of Duncan's multiple comparison test on the direct expansions of the number of expected farrowings in the second quarter.

Missing Data Procedure	Estimate : Estimate from "Reported" Procedure	Groupings
"Reported"	1.00	
Hot Deck	1.02	
Regression (BRR)	1.03	
Ratio (BRR)	1.03	
Hot Deck (BRR)	1.03	
Regression	1.04	
Ratio	1.05	
Complete	1.13	

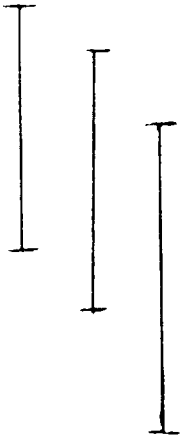


Figure 11: The results of Duncan's multiple comparison test on the estimated standard errors of the number of total hogs.

Missing Data Procedures	Estimate : Estimate from "Reported" Procedure	Groupings
Hot Deck	0.81	
Regression (BRR)	0.95	
Ratio (BRR)	0.96	
Hot Deck (BRR)	0.98	
"Reported"	1.00	
Ratio	1.05	
Regression	1.05	
"Complete"	0.97	

Figure 12: The results of Duncan's multiple comparison test on the estimated standard errors of the number of expected farrowings in the first quarter.

Missing Data Procedure	Estimate ÷ Estimate from "Reported" Procedure	Groupings
Regression (BRR)	0.78	
Ratio (BRR)	0.80	
Hot Deck	0.88	
"Reported"	1.00	
Hot Deck (BRR)	1.01	
Ratio	1.22	
Regression	1.22	
"Complete"	0.93	

Figure 13: The results of Duncan's multiple comparison test on the estimated standard errors of the number of expected farrowings in the first quarter.

Missing Data Procedure	Estimate ÷ Estimate from "Reported" Procedure	Groupings
Hot Deck	0.84	
"Reported"	1.00	
Ratio (BRR)	1.02	
Regression (BRR)	1.03	
Regression	1.16	
Ratio	1.22	
Hot Deck (BRR)	1.24	
"Complete"	0.96	

Table 13 displays the minimum difference needed to declare the estimates from two missing data procedures significantly different at a 5 % level. Thus, Table 13 is a useful tool in determining the power of Duncan's multiple comparison tests in this report. These minimum differences are from formula and are given as a percentage of the direct expansion from the "reported" procedure. One can see from Table 13 that Duncan's test is much less sensitive for the number of expected farrowings in the first quarter than for the other variables. One can also see that Duncan's tests are not very powerful for the estimated standard errors of any of three variables. This result highlights the extreme variability of the standard error estimates. However, the high variability is not too serious since estimated standard errors have less priority than direct expansion in this report. Still, more precision on the estimate standard errors is desirable.

Table 13: Minimum differences required for significance in Dunan's multiple comparison tests. Differences are a percentage of the direct expansion from the "reported" procedure.

Type of Estimate	Variable	Minimum Percentage Difference
Direct Expansion	Total Hogs	±0.033
	Pseud 1	±0.017
	Pseud 2	±0.014
	Expected Farrowings in the First Quarter	±0.111
	Expected Farrowings in the Second Quarter	±0.040
Estimated Standard Error	Total Hogs	±0.446
	Expected Farrowings in the First Quarter	±0.530
	Expected Farrowings in the Second Quarter	±0.404

To obtain a more powerful test of the differences in estimated standard errors this study uses multivariate t-tests. These multivariate tests attempt to distinquish between:

1: the hot deck procedure and the hot deck procedre (BRR)
2: the ratio procedure and the ratio procedure (BRR)
3: the regression procedure and the regression procedure (BRR)
4: the ratio porcedure (BRR) and the regression procedure (BRR)

The test statistic is the Hotelling-Lawley trace statistic which is defined to be:

$$T = \text{trace } (E^{-1} H)$$

where

E = error sums of squares matrix for the experimental design

H = Hypothesis sums of squares matrix for the experimental design.

For the four comparisons above the results are:

1: T = 7.15, significance level = 0.18
2: T = 5.28, significance level = 0.23
3: T = 4.85, significance level = 0.25
4: T = 0.44, significance level = 0.83

The first three tests are borderline cases. Because importance is attached to estimates of standard errors, one may accept larger signficiance levels as evidence of a difference in the procedures. However, the signficance levels of the first three tests are still not small enough to conclude that differences do exist. They are small enough to point out the need for further evidence. Other data sets must be analyzed to get more information about the procedures' effects on estimated standard errors. The power of the multivariate test can not overcome the high variability of these estimates.