ERROR PROFILE - MULTIPLE FRAME DESIGNS

Norman D. Beller, United States Department of Agriculture

I will introduce my topic by defining what I like to call a statistical paradox from a samplers point of view. One normally may assume that the primary purpose of sampling is to obtain needed information about the target population by measuring only a portion of the population due to costs, the destructive nature of sampling, or because population characteristics change rapidly. A sampling statistician's goal is to minimize variation within cost restraints in any survey, and probably more so in a repetitive survey program. Generally, the impact of this minimizing process produces a more complex survey design, questionnaire, and/or estimation procedure. These additional complications may create situations promoting increased nonsampling error. This, then, is the paradox: continued efforts to decrease sampling error (improve precision) often involve greater survey design complications which can increase the nonsampling error (decrease accuracy) and in turn, may result in a greater total error. There is some evidence to suggest that the surveys that I will discuss are victims of this paradox.

Multiple frame estimation implies the use of two or more sampling frames. This procedure allows greater coverage of the target population if no single complete frame exists. Multiple frame estimation provides greater efficiency if one can use less expensive data collection procedures on at least one of the frames. ESCS surveys generally use an incomplete list frame combined with a complete area frame. Efficiency is our major objective for using multiple frame methodology.

Multiple frame surveys are susceptible to errors inherent in each frame, plus errors which stem from associating the overlapping portions of the frames. These errors may individually have either a positive or negative effect upon the estimator and, as a result, may have either an additive or compensating effect. One must proceed carefully when implementing changes as a change may result in an estimate with a greater bias unless the nature of the errors is known.

The questionnaire has several concepts to develop in addition to collecting the required data in multiple frame survey methodology. Through the questionnaire, one must be able to associate the reporting and sampling units, provide information for overlap and nonoverlap determination, and possibly weights for computing a weighted nonoverlap estimator. The sample unit from the list domain is normally a name and address from the list, while the reporting unit from both the area and list frame is usually land operated and the livestock on that land at the time of the interview. To establish this association, the respondent is asked several questions to determine operated land.

A number of studies have pointed to difficulty in the use of questions relating to land for the purpose of associating the reporting unit and the sampling unit. In these same studies efforts were made to determine the net effect of editing to make data conform to survey concepts and/or for internal consistency. Resulting estimates when recalculated varied from the original estimates by 6 to 10 percent which is 2 to 4 times the magnitude of the sampling error.

Domain determination is one of the most critical procedures in multiple frame estimation. Since the area frame is a complete frame, the overlap between the two frames is identified by determining whether each reporting unit found in the area sample also could have been selected from the list frame. Overlap between the two frames is then determined by matching names associated with their respective reporting unit. This becomes extremely difficult with joint farming operations. Also, the use of nicknames, nonperson names, names primarily generated for legal purposes, and minimal address information all add to the difficulties of accurate matching via the use of names and addresses.

Our studies have provided ample evidence that there are nonsampling errors associated with domain determination. To date, the only methods to control this source of error have been attempts to obtain more complete name and address information, develop automated linkage procedures, and consider the size of the list frame sampled for multiple frame purposes. It is probably a safe assumption that the magnitude of errors arising from domain determination are positively correlated with the proportion of the universe operations covered by the list frame. If this is true, a relevant question becomes "How much of the universe should one attempt to cover with a list frame?"

Throughout the history of the multiple frame program the contribution to the sampling error and the resulting estimates attributable to the area nonoverlap have been larger than desirable. Generally the area nonoverlap domain has contributed about 20 percent of the estimate and 50-70 percent of its variability. Little if any success has been achieved in reducing contributions of the area nonoverlap in terms of either level or variability, regardless of the amount of effort placed on improving the list frame of all farms in terms of completeness. This phenomenon can only be explained by recognizing what is taking place in the area frame. As the list approaches completion and is sampled in its entirety, the item of interest becomes an increasingly rare item in the nonoverlap domain of the area frame. The area nonoverlap estimator becomes less efficient for fixed sample size as the item becomes rarer. Thus, the net result of increased resources being spent for list improvement, coupled with sampling the resulting list in its entirety, are largely negated by declining efficiency in the area nonoverlap domain.

Starting in 1974, a series of studies were conducted to determine the optimum mix of area and list frames. The objective of the analyses sought to determine if the size of the list could be reduced without seriously increasing the sampling error, thereby reducing the impact of nonsampling errors associated with domain determination. Analyses over several years for many different States all reached the same conclusion: it is not necessary to sample the entire list frame for the cattle and hog program given the current area sample. Significant reductions of list sample size and respondent burden can be realized and the size of list frame reduced substantially. Nonsampling errors associated with domain determination would be reduced.

The type of estimator used provides a partial answer to the nonsampling error problem. The screening estimator 1/ has been adopted in favor of the full multiple frame estimator 2/ in ESCS. The screening estimator is obtained by adding an area frame estimator for the nonoverlap domain (list incompleteness) to the estimate from the list frame for the overlap domain. This estimation procedure causes concern when one considers the errors arising from inaccurate domain determination. Conceptually the bias caused by improper domain determination is offset in the other frame. In other words, if the area frame nonoverlap estimate is biased downwards by classifying certain area frame respondents as overlap, when in truth they were not represented on the list frame, the area overlap estimate would be biased upwards. A full multiple frame estimator in this situation would be expected to reduce the impact of nonsampling errors arising from domain determination due to weighting the overlap domains together.

The weighted segment estimator is utilized for the nonoverlap domain in the current program. Entire farm data is weighted into the segment based upon the ratio of land area of the farm inside a segment to the land area of the entire farm. The condition required for the weighted estimate to be unbiased is that the sum of the weights for each population unit equals one. However, a biased estimate results if the data used to compute the weight is improperly reported. Generally, experience has shown that one of the more difficult items for farmers to report is the total land of the farming operation. Studies using reinterview techniques indicate that total farm acreage data ranged from 3 to 11 percent above the original survey indications resulting in a built in upward bias of that magnitude.

Another potential source of error arises from nonresponse. ESCS experience shows that the nonresponse problem is greater in the list frame compared with the area frame. The area frame nonresponse rate is between 2 and 10 percent, while the list frame nonresponse rate is substantially larger. Nonresponse and data imputation research has shown that there are feasible methods of reducing the relative bias caused by substituting respondent means for nonrespondent means. Nonetheless, all viable procedures rely on high quality historic data (control). The quality of the control data must be improved before any improved imputation procedures may be adopted. Meanwhile, an estimator has been developed that adjusts for the differing amounts of zero reports in the respondent and nonrespondent groups. Use of this estimator would reduce the relative bias and increase the list frame estimate.

In summary, many of the sources of errors that have been found arise from greater complications in the survey process. Many of the complications came about because of the desire for a lower sampling error without additional resources. The preceding material is a synopsis of a more detailed publication by the same title.

1/ Cochran, Robert S., "Multiple Frame Sample Surveys," ASA Meetings, University of Wyoming, December 1964.

2/ Hartley, H. O., "Multiple Frame Surveys," Proceedings of the Social Statistics Section, ASA (1962).