ON UNBIASED ESTIMATION IN THE AREA AND LIST FRAMES

Nicholas J. Ciancio and Robert D. Tortora, Statistical Reporting Service, USDA

INTRODUCTION

In a standard sampling text one defines a target population, specifies a sampling method, then caculates an estimate and variance. From a practical standpoint one must: (a) define the target population, (b) determine sampling frame units to reach the target population, and (c) define rules to associate the target population units with the sampling frame units.

Since the sampling units are the units that the sampling frame is divided into for sampling purposes they may not be the same as the target population of units. This affects (i) the sample design in that (a), (b), and (c) above have some impact on whether single or multiple frames are used and how samples are allocated between and within frames, (ii) the questionnaire design since the questions must determine which population unit is being associated with the sampling unit, (iii) the collection procedures since the sampling unit determines what is to be sampled, and (iv) the estimation procedures because they are directly related to the sample design, and should be unbiased if so desired.

Expected values are taken to determine if the estimators are unbiased. These estimators must remain unbiased under the survey operational conditions. This paper will show that under practical situations the estimates remain unbiased when different rules are used to associate the sampling unit and population unit under a simple random sample design.

This problem is discussed for two reasons. First, the Statistical Reporting Service, in its continued attempts to improve survey methodology, evaluates and compares rules to associate target population units with sampling frame units. Secondly, the problem of correctly using the frame to reflect the target population is critical to the success of any survey. The importance of this problem is noted by Cochran [1] and Deming [2] as discussants to a paper presented by Hansen, Hurwitz and Jabine, [3]. However, in [3] no proofs are presented for the surveys they mention. Therefore, this paper demonstrates that such proofs are feasible and that they are necessary to insure that good survey procedures are followed.

TERMINOLOGY

Before attempting to show that unbiased estimates are obtained, we must define the target population units, sampling frames, and describe the rules of association between the target population units and sampling frame units.

The target population is all farms in the 48 states. From [6] a farm consists of the area or areas of land under one operation or management including land owned and rented minus land rented to others on which there will be crops, livestock, poultry, or expected sales of agricultural products at some time during the calendar year.

In the target population there are different types of land operations. One is an individual land operation in which a single person is solely responsible for making management decisions for his business. A joint land operation is one operated by 2 or more persons, each of whom contributes some or all of the money, property, materials or labor to carry on a joint business. Each person participates in the management decisions and shares the profits or losses. Examples of joint arrangements are partnerships, corporations, and institutions or cooperatives. Finally, managed land is an operation whereby a person is paid to make the day-to-day decisions for the farm. The target population has been structured in two ways with regard to joint operation which is discussed under Rules 1 and 2.

There are two frames or partial frames, the area frame and the list frame, that when combined cover the target population used for multiple frame estimation. The area frame consists of all land area within the states. The area frame covers 100% of the population, therefore, the area frame is a complete frame. The land area is classified (stratified) according to land use in order to achieve homogeneity within strata. For the area frame the sampling unit is a small section of land called a segment. A segment is a piece of land with boundaries delineated on a map. Every parcel of land within a segment must be accounted for in the survey.

Within each segment sampled, all farms whose headquarters are within the segment boundaries, are interviewed. Every population unit (farm) is assigned to only one sampling unit (segment) even though pieces of land area associated with the population unit fall within many sampling units since each farm can have only one headquarters. Each sampling unit may contain more than one population unit or no population units.

Situations arise where it will be necessary to distinguish to which sampling unit a farm belongs. This is done by an approach which requires a 1-1 correspondence between farm operators and farms. The approach is needed because it is possible for more than one person to be accepted as the farm operator of a particular farm. For example, suppose two brothers operate a farm jointly and live in different houses. Unless proper rules are formulated this farm could easily have a chance of being sampled twice.

For individual operations the residence of the operator is usually defined as the headquarters. The following are examples of possible rules to help determine the operator of types of jointly operated farms.

(a) In an individual operation only the individual can be associated with the operation. For example, suppose Bob Smith is on the list and Bob Smith is an individual operator then if his name is selected he will report for the farm.

(b) In a joint operation there are three possible kinds of sampling units:

1. A joint operation name is on the list but none of the respective names of individuals

who comprise the joint operation are on the list. Therefore, if Smith Brothers is on the list and that sampling frame unit is selected, the farm operated for the Smith Brothers will be associated with the sampling unit.

2. The joint operations name is not on the list but at least one of the individuals who comprise the joint operation is on the list, then information concerning the joint operation will be reported by the individuals. If Sam Smith is on the list and is a partner in Smith Bros., then if he is selected he will report for Smith Bros. farm. If all partners report for Smith Bros. farm, duplication will result.

3. Both the joint operation name and the names of the individuals who comprise the joint operation are on the list. For example, suppose Smith Bros., Sam Smith and Bill Smith are on the list. Suppose Smith Bros. is comprised of Sam and Bill Smith. If the name Smith Bros. is selected it will report for Smith Bros. If either the name Sam Smith or Bill Smith is selected they will report for individual operations of their own, if any, but not for Smith Bros.

As can be seen the rules given above are not the only rules which can be used. Other rules of association between the sampling unit and population unit can be developed.

(i) When all partners live on the farm, the person who makes most of the decisions should be considered the operator,

(ii) If one partner lives on the farm, and the others live elsewhere, the one living on the farm should be considered the operator,

(iii) If all partners live on the farm, and appear to share equally in the management, the oldest should be considered the operator,

(iv) If none of the partners live on the farm, the oldest should be considered the operator,

(v) In father-son arrangements, accept the definitions of the respondent as to whether it is (1) a partnership, (2) two separate operations, or, (3) one operation with the father in charge and the 4-H or F.F.A. projects of the son merely a part of or incidental to the fathers' overall farming operations.

For corporations or institution type farms the person who makes the day-to-day decisions such as planting, harvesting, and marketing is considered the operator.

The list frame is a list of names of persons involved in farming operations and their corresponding addresses. In the list frame this address does not always correspond to the headquarters as in the area frame. The address in the list is the place where the person wants all correspondence to be mailed. Not all target population units are assigned to the list frame units, i.e. the list may be incomplete. Duplication may also occur within the list frame when one or more outside sources are combined to update and maintain the master list. Duplicated information can also be obtained for joint operations if each partner is on the list as one or a combination of names.

Due to this duplicated information on the list, rules must be developed to define and associate a target population unit with a sampling frame. Vogel [5] investigates problems in multiple frame applications using three different rules of association. Two of those rules will be used for the purpose of this paper to define the operator(s) of the population units.

The first rule we will use is the simpler of the two rules. The purpose of this rule is to eliminate any bias associated with determining if the operator of farm land is part of a joint operation. Rule 1 specifies that:

(a) The land operated singly by an individual name or in the name of a joint operation can only be associated with one frame unit.

(b) All joint operations in the target population will be handled using rules (i) thru (v) above.

The second set of rules primary purpose is to minimize the effect of partnership operations on the sampling errors. Rule 2 relies on some basic assumptions; viz.,

(a) Each partner in a partnership can report for the partnership operation whether contacted through the area or list sampling frames.

(b) Each partner can also report his individual operation if there is one.

(c) Each partner can correctly identify all of the other partners.

(d) Every partner that appears in the list frame will be identified.

The joint use of the two single frames is referred to as multiple frame sampling. Since the area frame is complete, every list sampling unit on the list frame can be mapped to a sampling unit in the area frame but not every name associated with the area frame is on the list frame. Thus multiple frame sampling presents us with the problem of determining the overlap between the sampling frames. It is necessary to determine which population units from the area frame could also have been obtained through the list frame. This determination is conducted by matching names. The problem is compounded when one or more names could be linked with the same operation.

UNBIASEDNESS

Now that all definitions, rules of association and assumptions have been stated, unbiased estimates for the population totals will be derived. These unbiased estimates naturally assume that all rules have been properly executed in the survey operations. We will consider estimators for the different mappings at the single frame level before combining frames. We assume simple random sampling.

The first frame to be considered is the area frame. The population units are farms as defined by Rule 1 and the sampling units are segments. As a means of identifying a farm with a unique segment we used the aforementioned headquarters rule. Let M = total number of population units (farms)

$$a^{X_i} = \begin{cases} \text{Number of farm headquarters in the i-th sampling unit (segment)} \\ 0 \text{ if the i-th sampling unit (segment) contains no farms.} \end{cases}$$

Since the area frame is complete we have $M = \sum_{i=1}^{a^N} a^{X_i}$

where $_aN$ is the total number of sampling units (segments). To estimate M we use $\hat{M}$ where

$$\hat{M} = \sum_{i=1}^{a^n} \frac{a^N}{a^n} \, {}_aX_i$$

$$= \sum_{i=1}^{a^N} \frac{a^N}{a^n} \, {}_aX_i \, \tau_i, \text{ where}$$

$_an$ is the number of segments sampled and

$$\tau_i = \begin{cases} 1 \text{ if i-th frame unit is selected} \\ 0 \text{ otherwise} \end{cases}$$

and $E(\tau_i) = \frac{a^n}{a^N}$

The expected value of $\hat{M}$ is clearly M.

The listed frame is a more complex problem than the area frame. In the list frame certain problems arise in mapping from the population units to the sampling units. The cases that will be considered are:

1. the list frame without joint operations or duplication under Rule 1.
2. the list frame with duplication under Rule 1.
3. the list frame with joint operations and duplication under Rule 2

Consider the list frame assuming it does not contain joint operations or duplication. The population unit is the farm and the sampling unit is a name and its corresponding address. Then there is a mapping from the target population units to the sampling frame units under Rule 1. Under Rule 1, which states that each name will report for itself, cases 1 and 2 can be shown to be unbiased.

Let $_LM$ = number of population units accounted for by the list frame.

Then $$_LM = \sum_{i=1}^{L^N} {}_LX_i,$$

where $$_LX_i = \begin{cases} 1 \text{ if name = farm} \\ 0 \text{ if name} \neq \text{farm} \end{cases}$$

and $_LN$ = total number of list frame units

Note: $_LM < M$ since the list is incomplete.

An estimate for $_LM$ based on a sample of size $_Ln$ is $\hat{{}_LM} = {}_LN \, \bar{{}_LX_i}$

$$= {}_LN \sum_{i=1}^{L^N} \frac{{}_LX_i}{L^n}$$

$$= {}_LN \sum_{i=1}^{L^N} \frac{{}_LX_i}{L^n} \, \tau_i$$

where τ_i is as defined before. Again $\hat{{}_LM}$ is an unbiased estimate of $_LM$.

The second case for which Rule 1 applies is the list frame with duplication. Rao [4] developed a procedure for handling duplication within the list where the number of times an operation can be selected is known. To develope the unbiasedness let

$_LN$ = total number of list frame units

$_LN'$ = total number of unique list frame units

$$_LA_i = \begin{cases} \text{number of times each unit is duplicated} \\ 1 \text{ if not duplicated, } i = 1, \ldots, {}_LN \end{cases}$$

$$_LX_i = \begin{cases} 1 \text{ if farm = name} \\ 0 \text{ if farm} \neq \text{name, } i = 1, \ldots, {}_LN \end{cases}$$

We then have

$$_LM = \sum_{i=1}^{L^N} \frac{{}_LX_i}{{}_LA_i} = \sum_{i=1}^{L^{N'}} \frac{{}_LX_i}{{}_LA_i} \, {}_LA_i = \sum_{i=1}^{L^N} {}_LX_i$$

$$\therefore \hat{{}_LM} = \sum_{i=1}^{L^n} \frac{L^N}{L^n} \frac{{}_LX_i}{{}_LA_i} = \sum_{i=1}^{L^N} \frac{L^N}{L^n} \frac{{}_LX_i}{{}_LA_i} \, \tau_i$$

where

$$\tau_i = \begin{cases} 1 \text{ if i-th frame unit is selected} \\ 0 \text{ if i-th frame unit is not selected,} \end{cases}$$

$i=1,\ldots, {}_LN$. Upon taking the expected value of $\hat{{}_LM}$ we obtain the desired unbiasedness of the estimator.

In contrast to Rule 1, Rule 2 allows for a population unit to be associated with more than one operator in the case of joint operations. Therefore, Rule 2 adds a new dimension to the problem of unbiased estimates by introducing another kind of duplication. As can be seen Rule 2 is more difficult to apply than Rule 1.

We will now consider the list frame with joint operations and duplications under Rule 2. Define $_LN$, $_LN'$ and $_LA_i$ as before and let

$$_LA'_i = \begin{cases} \text{number of times the population unit is} \\ \text{uniquely duplicated by different persons} \\ 1 \text{ otherwise, } i = 1, .., {}_LN \end{cases}$$

In determining $_LA'_i$ we are concerned with duplication of population units whereas $_LA_i$ was concerned with the duplication of frame units.

To see the use of $_LA_i$ and $_LA'_i$ consider the following situation for joint operations where the letters represent names with an address of persons who can report for some population unit:

Population	List Frame Units
1. A, B and C	1. B
2. D and E	2. A, B
	3. A
	4. C
	5. D
	6. A

The only duplication of frame units occurs between the third and sixth units. Therefore $_LA_i = 2$ for these two units and $_LA_i = 1$ for the remaining units in the frame.

$_LA'_i$ is determined by the number of times a population unit is uniquely duplicated by different persons. The first population unit is duplicated by persons corresponding to the first, second, third, and fourth and sixth frame units. But the third and sixth units are duplicates so they will be counted only once. This then leaves 4 unique frame units associated with the first population unit. We then associate an $_LA'_i = 4$ for the five frame units associated with the first population unit. Since only one frame unit is associated with the second population unit it

will have an ${}_{L}A'_{i} = 1$.

From the example above we can calculate ${}_{L}M$, where ${}_{L}X_{i} = 1$ for all frame units since each farm can be associated with a frame unit. Then

$$L^{M} = \sum_{i=1}^{6} \frac{L^{X_i}}{L^{A_i} L^{A'_i}} = 2$$

Therefore ${}_{L}M = 2$, which is the number of population units for this example. An estimate of ${}_{L}M$ is

$$\hat{L^{M}} = \sum_{i=1}^{L^{n}} \frac{L^{N}}{L^{n}} \frac{L^{X_i}}{L^{A_i} L^{A'_i}}$$, which is clearly unbiased assuming a properly executed mapping has occurred.

Up to this point we have discussed two rules of association between the population units and the sampling units for single frames, i.e., the area frame and the list frame. We would now like to mention the multiple frame.

The procedures used rely on certain assumptions in application of the two previously stated rules (see Rule 1 and 2). Decision diagrams for both sets of rules are generally used to determine the nonoverlap between frames. Since the area frame is complete and the list frame is incomplete the expression for M, the total number of population units, is

$M = {}_{NOL}M + {}_{OL}M$, where

NOL = not on list and OL = on list. From this expression it is easily seen that ${}_{OL}M = {}_{L}M$, the total number of population units associated with the list.

$$M = {}_{NOL}M + {}_{OL}M = {}_{NOL}M = {}_{L}M$$
$$= {}_{NOL}M + P\,{}_{OL}M + Q_{L}M, \text{ where } P + Q = 1.$$

The expression ${}_{NOL}M + P_{OL}M$ is associated with the area frame and $Q_{L}M$ is associated with the list frame.

SUMMARY

In this paper two rules for associating a target population unit with a sampling frame unit were presented. The first rule stated that each population unit reports for only its farm. The second rule stated that a frame unit can report for each population unit it is affiliated with. This is a drastic difference from the first rule in application. Much more work is needed in the form of checking the frame for the other members of joint operations and in the form of actually prorating the data. This additional work may lead to an increase in nonsampling errors for a survey.

The two rules were collectively applied to the area frame and the list frame under certain conditions. In both cases unbiased estimates of the total number of farms represented by each frame were obtained.

Additional rules should be developed and tested to associate the target population unit with the sampling frame unit. The importance of such considerations in surveys which required unbiased estimates is stressed.

REFERENCES

[1] Cochran, W.G., Discussion, International Institute of Statistics, 40 (1963), 538-539

[2] Deming, W.E., Discussion, International Institute of Statistics, 40 (1963), 543

[3] Hansen, M.H., Hurwitz, W.N., and Jabine, T.B., "THE USE OF IMPERFECT LISTS FOR PROBABILITY SAMPLING AT THE U.S. BUREAU OF CENSUS", International Institute of Statistics, 40 (1963), 497-517

[4] Rao, J.N.K., "SOME NONRESPONSE SAMPLING THEORY WHEN THE FRAME CONTAINS AN UNKNOWN AMOUNT OF DUPLICATION", Journal of the American Statistical Association, March 1968, 87-90

[5] Vogel, F.A., "SURVEYS WITH OVERLAPPING FRAMES - PROBLEMS IN APPLICATION", Proceedings of the Social Statistics Section, American Statistical Association, 1975

[6] "1976 June Enumerative and Multiple Frame Surveys - Interviewers' Manual' USDA