IS VALUE ADDED BY ADJUSTING FOR DUPLICATES IN A POPULATION FRAME?

Jason S. Lee, U.S. General Accounting Office
441 G. Street N.W., Washington, DC 20548

KEY WORDS: Sampling, Cost Efficiency

1. Introduction

In simple random sampling (SRS) designs it is assumed that all sample elements have a known, non-zero, equal probability of selection. Although unequal probabilities of selection can be an integral part of complex sampling strategies, statisticians warn that in SRS designs "[b]iases may arise where sample units are selected with known but varying probabilities" (Hansen, Hurwitz and Madow 1953, p. 59). They caution that "[a] frequent source of bias in sample designs is the use of varying probabilities in selecting the sampling units" (ibid.). In this paper I examine one form of departure from the equal selection probabilities assumption and I assess whether adjusting for this departure is an efficient use of resources.

Efficiency can be assessed as the amount of statistical accuracy gained for a given expenditure of resources. I assume that all survey efforts have budgetary constraints and that assorted error reduction schemes "compete" for available funds. At the present time there is not enough empirical data available to guide choices between competing approaches to error reduction. In response to this shortage, this paper presents evidence of little value added when SRS derived estimates are adjusted to account for unequal probabilities of selection. Limitations to the generalizability of this finding are discussed.

There are various ways unequal probabilities of selection may arise in SRS designs. If, for example, a sample of doctors is selected from patient lists, doctors with more patients have a greater likelihood of selection into the sample than doctors with fewer patients. Similarly, if families with children in school are the unit of analysis and are sampled from student enrollment rosters, then families with more than one child in school have greater selection probabilities. Or, if a sample of business owners is drawn from a list of businesses, people who own more than one business are more likely to be selected.

Another common source of unequal selection probabilities is the existence of duplicate listings in a population frame. This occurred in a recent study conducted by the U.S. General Accounting Office (GAO). By merging two somewhat overlapping lists, in order to obtain sufficient target population coverage, an unknown amount of duplicate listings was produced in the unified population frame.

Sometimes duplicate listings can be identified and purged before sample selection with the assistance of text-matching computer programs. However, in practice, the utility of this approach is often limited. An alternative to manually purging the entire population frame of duplicates--described in detail below--requires knowing how many times each sample element appears in the population frame.

Suppose you want to estimate the percentage of a population that would select a particular response option such as "yes" or "no," a degree of satisfaction from "very satisfied" to "very dissatisfied," or an age or sex or racial group identifier. To estimate the numerator (x'), score a variable, x_i, one if the response option of interest is selected and zero otherwise. The following formula estimates the total number in the population that would select this response option.

$$x' = \frac{N}{n} \sum_{i=1}^{n} \frac{x_i}{a_i}$$

The a_i is the number of times the *ith* sample element appears in the population frame.[1] Note that division by a_i is necessary to offset the increased likelihood that population elements with duplicate listings appear in the sample. To obtain the denominator (y') for the estimate of the percentage of a population with a particular attribute, you need to estimate the number of unique elements in the population. This estimate is given by the following formula:[2]

$$y' = \frac{N}{n} \sum_{i=1}^{n} \frac{1}{a_i} .$$

Here, division by a_i "corrects" N for duplicate listings. Now we can see that the ratio x'/y' is an unbiased estimate of the percentage of population elements that have the x_i attribute.

As noted, this procedure assumes that a_i is identified, which is to say that one must count the number of times each randomly selected sample element appears in the population frame.[3] This method of adjusting for duplicates may be far more practical than manually purging an extensive population frame. However, it remains to be seen whether in a situation like the GAO study it is an efficient use of resources to adjust for duplicates at all. In the following analysis, I assess the accuracy gained by using information from the sample to adjust for duplicates in the population frame. Then I evaluate the benefits received in light of the costs expended and discuss the findings.

2. Data

The General Accounting Office was asked by the Congress to determine what kinds of barriers nonprofit organizations face in acquiring federally-held foreclosed properties to assist the homeless. Although this job was scoped as a nationwide sample survey, we were not able to find a single list with adequate target population coverage. After reviewing multiple lists and considering various ways of combining them, we decided to merge the Federal Emergency Management Association (FEMA) list of nonprofit organizations with an independently constructed list maintained by the HOPE Foundation, a private, Texas-based organization. There were two sources of duplicates in this unified population frame of 14,018 entries. One was simply that a number of organizations were common to both lists. The other was due to the fact that FEMA adds an entry to its annual list each time it issues a check to an organization, and some organizations receive multiple checks in the same year. For the purposes of the GAO study, a simple random sample of 600 organizations was selected from the population frame.[4]

It took more than 320 person hours (more than eight weeks) to count the duplicates in this sample of 600. On average, this is about 1.9 hours per sample element. Discriminating among organizations often proved to be an onerous task. Each entry had numerous fields of information. Sometimes the organization names and addresses and phone numbers were the same but the contact persons differed. Sometimes the contact persons were the same but the organization names and addresses and phone numbers were different. Compounding the task further still, some organizations that initially looked the same were actually different, perhaps because a local Red Cross or Salvation Army (which appeared as the primary address) acted as a fiscal agent for multiple smaller concerns (listed as secondary addresses), or perhaps because several independent organizations had banded together under an umbrella agency. All in all, there were many combinations of same and different information. In the end, decision rules were applied, but in many instances the organizations in question had to be contacted by telephone. Once the counting was completed, we had learned that of the 600 sampled organizations, 200 were duplicated at least once in the population frame. Table 1 shows the distribution of sample members by the number of entries in the population frame.

3. Results

I used a ratio estimation program (written and reviewed within GAO and referenced to formulas in Hansen, Hurwitz and Madow (1953) and Cochran (1977)) to calculate ratio estimates (total number with an attribute[5] divided by total number in the population, i.e., x'/y') and variances of these estimates. As described earlier, the variables in these ratios are weighted by $1/a_i$ to adjust for duplicates in the population frame. These estimates and their sampling errors at the 95% confidence level are listed in columns 1 and 2 of Table 2.

How much accuracy would have been lost had the estimates not been weighted by $1/a_i$ to adjust for duplicates? In columns 3 and 4 of Table 2, I present the estimates of the percentage of organizations with a given attribute, and their sampling errors calculated at the 95% confidence level, derived *without* adjusting for duplicates in the population frame. Because the adjusted percentages (column 1 of Table 2) are unbiased estimates of the population values, they are the standard against which the unadjusted estimates (column 3) are compared. Of the 23 possible comparisons, all of the unadjusted estimates are within the 95% confidence intervals of the adjusted estimates. Moreover, about two thirds (15 of 23) of the differences between the adjusted and unadjusted estimates are less than one percentage point. (In all, about half (12) of the unadjusted estimates are a little larger and about half (11) are a little smaller than their adjusted counterparts.) Insofar as the precision of these estimates is concerned, in all but one case the variance of the unadjusted estimates is slightly smaller than or is equal to the variance of the adjusted estimates. Thus failing to adjust for duplicates in the population frame did not result in significantly less precise estimates of population percentages.

4. Discussion

In this paper I assessed the costs and benefits of adjusting sample estimates for duplicate listings in a population frame. The data I used are from a General Accounting Office study of homelessness assistance nonprofit organizations and the barriers they experience to acquiring federally-held foreclosed property. Primarily because it was hard to determine whether some of the sample organizations were the same as or different from other organizations in the population frame, the cost of weighting the data to adjust for unequal selection probabilities was high. More than eight weeks of person hours were expended to count and validate the number of times each sample member appeared in the population. Moreover, as compared to unadjusted SRS designs, additional resources were used to calculate ratio estimates and their variances. These costs can be measured in hours or dollars or, alternatively, as foregone opportunity costs. To decide whether or not to adjust for duplicates, the survey specialist must judge whether the accuracy gained is worth the costs expended, or whether it is better to subtract these costs or perhaps to use them to reduce nonsampling sources of error (see, for example, Groves, 1989).

This GAO study is a situation in which failing to adjust for duplicates in the population frame resulted in percentage estimates that were not substantially different from adjusted (unbiased) estimates. Moreover, the unadjusted estimates are at least as precise as the adjusted estimates. These findings support the conclusion that the benefits from adjusting for duplicates in the population frame were not worth the costs expended.

Is this conclusion generalizable to other survey situations? There are a number of factors to consider when a simple random sample design contains duplicate listings in the population frame. First, the whole problem may be dismissed if purging a population frame of duplicates is a relatively simple matter.

If purging duplicates is not easily accomplished, then one must assess whether adjusting for duplicates is likely to be an efficient expenditure of available resources. There are at least three important considerations. The first is straightforward and depends on the type of estimates needed from the data. In this study I focused only on percentages. If population counts or total expenditures or any other aggregate estimates of numbers in the population are required, there is no comparable alternative to counting the sample duplicates.

The second and third considerations require some degree of conjecture (perhaps supplemented by exploratory data analysis). Using any available information, consider the potential *volume* of multiple listings. In the GAO study, sample estimates indicate that about one third of the population had multiple listings, though most of these were single duplicates. If, in a different situation, a larger portion of the population frame contains duplicates, or if duplicated elements tend to be listed more than twice, then unweighted estimates may be more biased than the ones reported in this paper.

Finally, carefully consider the *underlying process* that distributes duplicates in the population frame. Adjusting for duplicates is less critical if the process is random with respect to important study variables than if it is systematic or ordered. The analysis reported in this paper suggests that in the GAO study the duplicate listings were randomly distributed with respect to all of the variables examined. Even a considerable departure from the equal selection probabilities assumption of SRS designs introduced insignificant bias and no loss of precision. Presumably, this is due to the randomness of the distribution of duplicates in the unified population frame. However, if the process that distributes duplicates in a population frame is ordered or systematic with respect to important study variables, then assume that the cost of adjusting for duplicates is worth the benefit gained by virtue of significantly less biased estimates. Hansen, Hurwitz and Madow (1953) gave sound advice. They warned that biases may arise and cautioned of the need for care in dealing with samples selected with known but varying probabilities.

Note: The views expressed in this paper are those of the author and do not necessarily reflect the position of the U.S. General Accounting Office.

Table 1. Number of Times Sample Members Appear in Population Frame

Number of entries in the population frame	Number of sample members	Percent of respondents
1	400	60.1%
2	141	27.2
3	36	6.8
4	10	2.3
5	4	0.8
6	3	0.8
7	2	0.3
8	2	0.8
9	1	0.3
12	1	0.3
	600	383

References

Groves, R. (1989). *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

Hansen, M., Hurwitz, W., and Madow, W. (1953). *Sample Survey Methods and Theory. Volume 1: Methods and Applications.* New York: John Wiley & Sons.

Cochran, W. (1977). *Sampling Techniques (3rd ed.).* New York: John Wiley & Sons.

Notes

[1] See formula 4.1 (p. 61) in Hansen, Hurwitz and Madow, 1953.

[2] See Hansen, Hurwitz and Madow (1953, p. 63).

[3] If a duplicated population element is selected into the sample more than once (say, *x* times), its' records should be included in the data file *x* times.

[4] Of the six hundred organizations, five had single duplicates within the sample. These duplicates were allowed to remain in the sample when the adjusted estimates were calculated. (See footnote 2.)

[5] These attributes include: how much need there is in the organizations' service area for properties to be used to assist the homeless, how much interest nonprofit organizations have and what kinds of barriers they experience in trying to acquire foreclosed properties, what kinds of properties would they find useful (multiunit or single family), and in what federal homeless assistance programs they have participated.

Table 2. Estimates and Sampling Errors at the 95% Confidence Level, Adjusted and Unadjusted for Duplicates

	Adjusted for duplicates		Not adjusted for duplicates	
	% with attribute	95% sampling error	% with attribute	95% sampling error
	(1)	(2)	(3)	(4)
Need for Property:				
Very large	38.1	5.1	38.4	4.8
Large	32.0	4.8	33.9	4.7
Medium	20.1	4.2	19.5	3.9
Small	4.9	2.4	4.2	1.9
Very small	3.0	1.9	2.4	1.5
None	0.7	0.9	0.5	0.7
Missing	1.2	1.2	1.1	1.0
Barriers (% yes):				
Not enough info.	65.9	4.9	65.0	4.7
High rehab. cost	51.3	5.2	54.5	4.9
Neighbors opposed	26.3	4.6	26.1	4.3
Fed. $ not avail.	50.4	5.2	52.9	4.9
Most useful type of property:				
Single family house	17.1	3.9	18.7	3.8
1 unit in a bldg.	2.3	1.5	2.9	1.7
Duplex	9.0	3.0	9.2	2.8
Triplex	4.4	2.2	3.7	1.9
4 units	12.1	3.3	13.2	3.3
More than 4 units	20.4	4.2	21.1	4.0
Hotel or motel	17.5	4.1	15.5	3.6
Other	11.5	3.4	11.1	3.1
Participation in other Fed. programs:				
Emergency Shelter	52.9	5.2	56.3	4.9
Section 8 Rental	21.4	4.2	23.4	4.2
Section 8 SRO	5.3	2.4	5.0	2.1
Fed. Surplus Prop.	11.2	3.2	12.1	3.2