ESTIMATION WITH DUPLICATE SAMPLE UNITS FROM STRATIFIED SAMPLES

Iris Shimizu, National Center for Health Statistics

I. Introduction

In this paper we consider several estimators for totals when units are replicated in a sample, particularly stratified samples. Unless we state otherwise, we assume that the number of listings in the frame for individual units cannot be determined.

The problem considered here arose in connection with the 1973-74 Nursing Home Survey which uses a stratified sample. The sampling frame consists of all nursing homes listed in the Master Facility Inventory maintained by the National Center for Health Statistics and the homes listed by the Agency Reporting Service as having come into existence since the last census of homes that was conducted for the Inventory in 1971. Efforts were made to "unreplicate" the frame by matching names and addresses of the homes listed.

Despite the efforts to remove replication from the sampling frame, 34 homes were included in the sample twice. These were only discovered by interviewers who found themselves going back to some of the homes they had already interviewed. While the replicated sample units appeared exactly twice, each, in our sample, it is possible that a unit could be listed three or more times in the frame. It is not feasible to determine the exact number of times a unit is listed in the frame, either from the frame or from the unit itself. Furthermore, the sample duplicates occurred across strata lines, rather than within strata. Strata were defined on the basis of bed size and whether the home was certified for Medicare and/or Medicaid.

The literature seems to contain only a few articles that deal with estimation in the presence of sample replicates. Even then, most authors such as Hartley (1962), Kish (1965), and Rao (1968) assume that the number of listings in the frame for each sample unit can be determined. Knowledge of replication in the frame is assumed unknown by Gurney and Gonzalez (1972) who compare the merits of several estimators that can be applied when replication occurs in a simple random sample. Their study favors an estimator in which each observation in the sample is weighted inversely by the number of times the unit is included in the sample. We include an extension of that estimator in our study.

We also consider an adaption of an estimator given by Simmons (1963) for use in the first cycle of the Nursing Home Survey. For the first cycle, a check of the frame was made to determine all the replicates in the frame for each sample case. Then a review of available records, forms, and correspondence determined which in a set of replicates was the "real" unit and which were "ghosts" replicates, i.e., that would be eliminated if all replicates in the frame were discovered. If the review failed, a unit was selected randomly from the set to be the "real" replicate. Simmons' unbiased estimator for unduplicated total measures for the i^{th} stratum was then formulated as

$$X'_i = \frac{M_i}{m_i} \sum_j B_{ij} X_{ij} + \left[\frac{M_i}{m_i} \sum_j (1-B_{ij}) \frac{W_{ij}}{A_{ij}} X'_{ij}\right] + \left[\sum_j (1-B_{ij}) \frac{W_{ij}}{A_{ij}} X'_{ij} \sum_\alpha \frac{M_\alpha}{m_\alpha} \sum_\beta {}_{ij}d_{\alpha\beta}\right]$$

where

M_i = total number of units listed in the frame for stratum i

m_i = number of sample units selected from stratum i

$B_{ij} = \begin{cases} 1 \text{ if } ij^{th} \text{ unit is } \underline{\text{not}} \text{ replicated in the frame} \\ 0 \text{ otherwise} \end{cases}$

x_{ij} = measure for the ij^{th} unit

x'_{ij} = measure for the "real" replicate of the ij^{th} unit

$W_{ij} = \begin{cases} 1 \text{ if the } ij^{th} \text{ unit is the "real" replicate} \\ 0 \text{ otherwise} \end{cases}$

A_{ij} = total number of times that the ij^{th} unit is listed in the sample frame

${}_{ij}d_{\alpha\beta} = \begin{cases} 1 \text{ if } \alpha\beta^{th} \text{ unit is a replicate of the } ij^{th} \text{ unit} \\ 0 \text{ otherwise} \end{cases}$

(Note that a "real" unit and a "ghost" unit may have different numerical values, even though they are replicates. In the Nursing Home Survey, the bed sizes recorded in the Master Facility Inventory for some replicated sample homes were different.) The bracketed quantities in the above equation are an unbiased estimate of the unduplicated component due to replicates that belong in the i^{th} stratum. The first bracket is the consequence of the ij^{th} unit's falling into the sample; the second is the consequence that the measure for the ij^{th} unit is used with a replicate which is also included in the sample.

II. Alternate Estimators

Here, we consider the merits of three estimators for stratified samples assuming only that the number of times a unit is replicated in the sample is known. Let

a_{ij} = number of times that the ij^{th} sample unit is replicated in the sample

$b_{ij} = \begin{cases} 1 \text{ if the } ij^{th} \text{ sample unit is } \underline{\text{not}} \text{ replicated in the sample} \\ 0 \text{ otherwise} \end{cases}$

Estimators for population totals may be written in the form

$$x'(k) = \sum_i x'_i(k)$$

where $x_i'(k)$ is the k^{th} estimator of the unduplicated stratum total for the i^{th} stratum. In the following we present only formulas for the different $x_i'(k)$ but we discuss the merits of the resulting estimators $x'(k)$ for unduplicated total measures.

Illustrations of how the estimators may behave are made by use of the toy model:

Stratum A	Stratum B
$A_1 = 1$	$B_1 = 3$
$A_2 = 3$	$B_2 = 4$
$A_3 = 5$	$B_3 = 7$

A_2 is the ghost of B_1.

Tables of estimates for samples from this toy model are given at the end of this paper.

1. Simple weighted estimator

Adapting the estimator favored by Gurney and Gonzalez (1972) to stratified sampling, the estimator for the i^{th} stratum becomes

$$x_i'(1) = \frac{M_i}{m_i} \sum_j \frac{X_{ij}}{a_{ij}}$$

If the population values A_{ij} were used in place of the sample values a_{ij}, $x'(1)$ for population totals would be unbiased. However, since the sample number of replicates will always be less than or equal to the true number of times that a unit is replicated in the frame, $x'(1)$ is biased in the positive direction. In practice this bias is expected to be small since the number of replicates in the frame will probably be minimal with respect to the size of the frame when efforts are made to unduplicate the frame.

Estimators of the individual strata totals may be biased in either direction, depending on the distribution of "ghost" units in the stratified frame. This may be observed in Table 1 for our example.

The second estimator considered here was derived after observing how $x'(1)$ behaved in some examples. In the toy model, if the replicated unit appears only once in the sample (e.g., Sample 2 in Table 1), the contribution to the estimate due to that single appearance is greater than that made by individual appearances of the unit when the unit is replicated in the sample (e.g., Sample 1). If the total of this excess in each stratum over all possible samples is divided by the total number of possible samples in which the unit may be replicated and then the results subtracted from the stratum estimator whenever the unit is replicated in the sample, the new estimator $x'(2)$ should be unbiased.

2. An unbiased estimator

The estimator can be written as

$$x_i'(2) = \frac{M_i}{m_i}\left[\sum_j \frac{X_{ij}}{a_{ij}} - \sum_j (1-b_{ij})\frac{X_{ij}}{a_{ij}} F_{ij}\right]$$

where F_{ij} is the ratio of the number of samples in which the ij^{th} unit can appear without replication over the number of samples in which the ij^{th} unit can be duplicated. The formula for F_{ij} depends on the maximum number of replicates occurring in the sample for the ij^{th} unit and the distribution of these replicates. For example, when units are replicated only within the same strata:

If the maximum number of sample replicates is 2, then

$$F_{ij} = \binom{M_i - 2}{m_i - 1} \Big/ \binom{M_i - 2}{m_i - 2}$$

or if the maximum number of sample replicates is 3, then

$$F_{ij} = b_{ij}(2)\binom{M_i - 2}{m_i - 1}\Big/\binom{M_i - 2}{m_i - 2} + 2b_{ij}(3)\left\{\binom{M_i-3}{m_i-1}\left(-\frac{2}{3}\right) + 3\binom{M_i-3}{m_i-2}\left[\binom{M_i-2}{m_i-1}\Big/\binom{M_i-2}{m_i-2} - \frac{1}{3}\right]\right\}$$

to allow for the fact that the unit may appear once or twice as well as three times. Here

$$b_{ij}(k) = \begin{cases} 1 \text{ if the } ij^{th} \text{ unit is included in the sample } k \text{ times} \\ 0 \text{ otherwise.} \end{cases}$$

Similar formula can be derived for cases in which a unit is replicated in more than one stratum. If $a_{ij} = 2$ and these replicates appear in strata 1 and 2, say, then

$$F_{1j} = \binom{M_1 - 1}{m_1 - 1}\binom{M_2 - 1}{m_2}\Big/\binom{M_1 - 1}{m_1 - 1}\binom{M_2 - 1}{m_2 - 1}$$

Similarly, F_{2j} is defined by replacing the 1's with 2's and vice versa.

The estimator is complex but it can be proven that $x'(2)$ for the unduplicated population total is indeed unbiased. However, it can also be proven that the mean squared error (MSE) of $x'(2)$ is greater than that of $x'(1)$. Proofs are left to the reader. Hence, $x'(2)$ does not appear to be a desirable estimator.

3. An estimator adjusted for replication across strata

The use of the sample value a_{ij} in place of the population value A_{ij} in Simmons' estimator leads to the estimator

$$x'_i(3) = \frac{M_i}{m_i} \sum_j b_{ij} X_{ij} + \frac{M_i}{m_i} \sum_j (1-b_{ij}) W_{ij} \frac{X'_{ij}}{a_{ij}}$$

$$+ \sum_j (1-b_{ij}) W_{ij} \frac{X'_{ij}}{a_{ij}} \sum_\alpha \frac{M_\alpha}{m_\alpha} \sum_\beta {}_{ij}d_{\alpha\beta}$$

In the sum over strata, the total weight given each sample unit by this estimator is the same as that given by x'(1), hence x'(3) gives the same estimates for the total population as does x'(1). That means that x'(3) is biased in the positive direction, as is x'(1), and that x'(3) has the same MSE as x'(1).

However, in x'(3) the total weight is given to only the "real" unit in each set of sample replicates. That is when the ij^{th} unit is duplicated in the sample, the weight of the "real" duplicate is $\frac{1}{2}[(M_i/m_i) + (M_\alpha/m_\alpha)]$, where the ij^{th} unit is duplicated in the α^{th} stratum. The weight given to the "ghost" is zero. Intuitively, estimators $x'_i(3)$ for unduplicated strata totals should have less bias than $x'_i(1)$. Table 3 displays this phenomena for the toy model but the theory has not been proven at this writing. It is clear that the sum of biases for the strata totals must equal the bias of the estimator for the total, regardless of the strata in which the bias occurs.

III. Conclusion

Three estimators for unreplicated totals are presented for use with samples in which replicates may occur. The unbiased estimator for population totals that we consider is undesirable because it is complex and because its MSE is greater than that of the two biased estimators also considered. While the other two estimators are biased, in practice the bias should be small when the amount of replication is small relative to the total population. The two biased estimators, in which sample measures are weighted inversely by the number of sample replications for each unit, give identical estimates for the totals and hence identical bias and MSE's for estimates of the total. The second of these estimators shifts all the weight from "ghost" replicates to the "real" replicates within sets of sample replicates. This suggests that while there is no difference between estimates of the two biased estimates for population totals, the estimator of stratum totals provided by the second of these could be less biased than that provided by the first estimator.

TABLE 1: Estimates for Toy Model Produced by x'(1) and Component Contributions to Estimates by Sample

Sample	Estimate of Total x'(1)	Contributions to Estimate					
		by stratum		by "reals" in stratum		by "ghosts" in stratum	
		$x'_A(1)$	$x'_B(1)$	A	B	A	B
1. $A_1A_2B_1$	8.25	3.75	4.5	1.5	4.5	2.25	-
2. $A_1A_2B_2$	18.0	6.0	12.0	1.5	12.0	4.5	-
3. $A_1A_2B_3$	27.0	6.0	21.0	1.5	21.0	4.5	-
4. $A_1A_3B_1$	18.0	9.0	9.0	9.0	9.0	-	-
5. $A_1A_3B_2$	21.0	9.0	12.0	9.0	12.0	-	-
6. $A_1A_3B_3$	30.0	9.0	21.0	9.0	21.0	-	-
7. $A_2A_3B_1$	14.25	9.75	4.5	7.5	4.5	2.25	-
8. $A_2A_3B_2$	24.0	12.0	12.0	7.5	12.0	4.5	-
9. $A_2A_3B_3$	33.0	12.0	21.0	7.5	21.0	4.5	-
Expected Value	21.5	8.5	13.0	6.0	13.0	2.5	
True Value	20.0	6.0	14.0				
Bias	1.5	2.5	-1.0				

TABLE 2: Estimates of the Unduplicated Total by Sample

Sample	Estimator		
	x'(1)	x'(2)	x'(3)
1. $A_1A_2B_1$	8.25	1.5	8.25
2. $A_1A_2B_2$	18.0	18.0	18.0
3. $A_1A_2B_3$	27.0	27.0	27.0
4. $A_1A_3B_1$	18.0	18.0	18.0
5. $A_1A_3B_2$	21.0	21.0	21.0
6. $A_1A_3B_3$	30.0	30.0	30.0
7. $A_2A_3B_1$	14.25	7.5	14.25
8. $A_2A_3B_2$	24.0	24.0	24.0
9. $A_2A_3B_3$	33.0	33.0	33.0
Expected Value	21.5	20.0	21.5
Bias	1.5	0	1.5
Variance	54.875	93.5	54.875
Mean Squared Error	57.125	93.5	57.125

TABLE 3: Expected Values and Biases of Estimators for Strata Totals

Estimators	Expected Value Stratum		Bias Stratum	
	A	B	A	B
x'(1)	8.5	13.0	2.5	-1.0
x'(3)	8.0	13.5	2.0	- .5

REFERENCES

Gurney, Margaret and Gonzalez, Maria Elena (1972). "Estimates for Samples from Frames Where Some Units Have Multiple Listings." *American Statistical Association Proceedings, Social Statistics Section.*

Hartley, H.O. (1962). "Multiple Frame Surveys." *American Statistical Association Proceedings, Social Statistics Section*

Kish, Leslie (1965). *Survey Sampling.* John Wiley and Sons, Inc.

Simmons, Walt R. (1963). "Duplication in the MFL--Amplification of memo of June 20, 1963, on Handling of Estimation in the MFL and RPS-I." Unpublished memo, National Center for Health Statistics.

Rao, J.N.K. (1968). "Some Nonresponse Sampling Theory When the Frame Contains an Unknown Amount of Duplication." *Journal of the American Statistical Association,* v. 63.