ESTIMATES FOR SAMPLES FROM FRAMES WHERE SOME UNITS HAVE MULTIPLE LISTINGS

Margaret Gurney and Maria Elena Gonzalez, U.S. Bureau of the Census

A. Introduction

It is often necessary to sample from a frame which is known to contain multiple listings for the same unit; this is particularly true for mail surveys. The estimation of totals and other statistics from such a sample presents some interesting and difficult problems.

In order to relate the subject matter of this paper to a real situation, we base our discussion chiefly on the sampling of small farms which was done as part of the 1969 Census of Agriculture. Units on the Census list which were expected to have Total Value of Products sold (TVP) less than $2,500 in 1969 were sampled at a rate of 1 in 2, and were sent a short questionnaire (form A2). Not all units on the mailing list represented farms according to the Census definition. In the interest of obtaining adequate coverage of farms, it was necessary to use several sources, not restricted to known "in-scope" farms, to construct the Census list. These lists were merged after unduplication on the basis of identical Social Security (SS) and Employer Identification (EI) numbers.

If anyone received more than one questionnaire, he was instructed to fill in any one of them, to mark the others "extra copy," and to return all of them in the same envelope. If any one of the forms returned by a respondent was an "A1," which was the regular form and was not subject to sampling, the form returned was assigned a weight of "1." If all of the questionnaires returned by a respondent were "A2's," the question of the best weighting procedure arose. Various estimation formulas can be used. Several are discussed below, including one based on a mathematical model for handling cases in which the respondent does not follow instructions.

B. Estimates for small farms when the respondent follows instructions

The properties of estimates discussed in this section depend on the following assumptions:

1. Respondents follow instructions, and return together all forms received by them.

2. The number of times a given unit is listed is uncorrelated with the characteristic being estimated. (For estimates of total number of units assumption 2 is not needed.)

Under these conditions we can make estimates which are unbiased, but have large sampling errors; estimates which are biased, but have smaller mean square errors, are also considered.

1. An unbiased estimate. An unbiased estimate, which is easy to apply, and which does not require that we know the total number of times a farm is on the listing, can be made by tabulating the data only for those farms for which the farm operator returned an odd number (1, 3, 5, ...) of questionnaires. These data are then multiplied by 2. This estimate is unbiased, which results from a property of the binomial expansion of $(1+1)^n$. To illustrate, consider farms which are on the list twice: approximately $\frac{1}{4}$ of them will not be selected, $\frac{1}{2}$ will be selected once, and $\frac{1}{4}$ will be selected twice. If we use only the farms which were selected once, and multiply the sum by 2, we shall have an unbiased estimate of the total of a characteristic for farms which are on the list twice. If we consider farms which are listed three times, the proportion which will not be selected at all is about 1/8; about 3/8 will be selected once; 3/8 will be selected twice; and 1/8 will be selected three times. The farms which are selected an odd number of times (3/8 plus 1/8 = 1/2) are used in making the estimate; multiplication by 2 will lead to unbiased estimates. The extension to higher amounts of duplication can be proved by induction.

This type of estimate is not a desirable one, since it is not using all of the data available: all of the farms which were selected an even number of times have a weight equal to zero. Consequently, the variance of this estimate may be larger than that of some other estimate.

2. An unbiased estimate for a list where every farm is listed k times. We consider, as the next step toward a better estimate, a list on which every listing is replicated exactly the same number of times, and ask for the optimum weighting scheme. If k is the number of replications, and N is the unduplicated number of farms, then the list contains $k \cdot N$ listings. If a sample of $\frac{1}{2}$ of the listings is selected, then approximately $[\binom{k}{1}/2^k]N$ will be selected once; $[\binom{k}{t}/2^k]N$ will be selected t times; etc. It can be shown that an unbiased estimate with minimum variance is obtained if all farms in the sample are given the same weight, and that this weight should be $2^k/(2^k-1)$. As k becomes large the optimum weight approaches unity.

Unfortunately, we do not know, when a farm operator returns one, two, or more copies of the A2 form, how many times his place was on the mailing list. We may have an approximate distribution, such as 80% of the farms are listed once, 10% are listed twice, 6% are listed three times, etc. But this doesn't help us to know how many times this particular farm was listed. Hence we go to another estimate, which is biased, but which can be expected to have a smaller mean square error for most items than the unbiased estimate described in section 1, above.

3. A reasonable biased estimate. Since the matching is presumed to have been reasonably successful, perhaps less than 20% of the farms are on

the listing more than once. This means that something more than 80% of the farms which respond with one questionnaire are actually singles, and should have a weight equal to 2. The remaining "single" responders should have weights of 4/3, 8/7, 16/15, etc., but we don't know which weight. Any farm which returned two or more questionnaires is on the list more than once, and with a sample of 1 in 2 has a high probability of being selected; hence it is reasonable to assign a weight = 1 to the "multiple" responders.

Since it is not possible to distinguish the "single" responders who are on the list more than once from those which are listed exactly once, we assign the weight "2" to them. The assignment of this weight to something less than 20% of the farms reporting once (the ones which are actually multiples) introduces an upward bias into the sample estimates. The size and importance of the bias will depend, of course, on the distribution on the list of single listings, doubles, triples, etc.; it will depend also on differences in the characteristics for farms which are listed once, as compared with farms which are listed more than once.

4. Illustration. Suppose that in a certain area there are 3200 farms, but some of them are on the mailing list more than once. Suppose the distribution by the number of replications is

Number of Replications	Frequency (P_i)	Number of Farms (N_i)	Number of Listings
1	.80	2560	2560
2	.10	320	640
3	.06	192	576
4	.03	96	384
5	.01	32	160
Total	1.00	3200	4320

The number on the listing is about 36% larger than the actual number of farms.

Now take a 50% sample of the listings; 2160 listings will be selected representing 1600 farms. The number of farms in the sample is a random variable, and will not be exactly equal to 1600.

Number of replications	Expected Distribution of the Sample: Not selected 0	Number of times selected: 1	2	3	4	5	Number of farms selected
1	1280	1280					1280
2	80	160	80				240
3	24	72	72	24			168
4	6	24	36	24	6		90
5	1	5	10	10	5	1	31
Total	1391	1541	198	58	11	1	1809

If we use the unbiased estimate from section 1. above, the expected value of the estimate of the number of farms is 2(1541 + 58 + 1) = 3200.

The biased estimate of section 3. leads to an estimated number of farms: n' = 2(1541) + 1(198 + 58 + 11 + 1) = 3350, which is an overestimate by 150 farms, or about 4.8%. This can be shown to have a smaller MSE than the unbiased estimate.

5. Another biased estimate. If we have any information which will allow us to estimate the approximate magnitude of P_1, the proportion of unduplicated units, it is possible to obtain a closer estimate of the number of farms, which will usually be an underestimate (depending largely on how well we estimate P_1). Notice that the proportion of farms which look like "singles" (1541/1809 = .852) is considerably larger than the true proportion of singles in the population, which is .80. If the respondent fails to follow instructions and returns only one form when he receives several, or returns more than one but in different envelopes, the proportion which look like "singles" will be even greater than .85. Consequently, it may be desirable to assign a weight of "1" to some small proportion of the apparent singles. A rough guess of the proportion of true singles can lead to a smaller mean square error; for example, if we guess the proportion of singles which are true "singles" to be .84, then the estimated number of farms is

$$n'' = 2(1541 \times .84) + 1(1541 \times .16 + 198 + 58 + 11 + 1) = 2(1294) + 1(515) = 3103$$

which is closer to 3200, the number of farms in the area, than the estimate of 3350, in section 3. In this illustration, a guess as low as 81%, or as high as 99% will lead to a smaller bias in the estimate of the number of farms than the simple procedure of section 3.

The discussion of the preceding sections can be easily extended to other sampling fractions. If the sampling fraction is less than $\frac{1}{2}$, the variance of the unbiased estimate (section 1.) is usually even larger, relative to the biased estimate of the latter part of section 3., and would not normally be considered a candidate for practical use.

C. Estimates when the respondent does not follow instructions

Respondents in a mail survey may fail to follow instructions in a number of ways. A respondent receiving t agricultural questionnaires may return any number from 0 to t of them. We have no way of knowing how many questionnaires a respondent received. Since the questionnaires he does not return are not known to be duplicates, they will probably be considered "nonresponses," and further followup will be done, and imputation will be applied to them if the followup is not successful.

It is possible to construct models which take into account the sampling fraction (f) and the probability that a respondent who receives t questionnaires returns r of them. Such a set of probabilities ($p_{t,r}$) can be applied to various sets of probabilities of multiple listings (P_t) to get some notion of the efficiency of various estimators. The models developed below start (section 1.) with a sampling frame consisting of k lists, each containing all N elements in the population (see section B2., above). Two methods are compared (section 2.): the method of section B1., and the latter part of section B3. The extension is then made (section 3.) to variable numbers of duplications.

1. Model[1/]

Let a population consist of N elements, and let the sampling frame consist of k lists, each containing all N elements. In each list, each element is given a probability f of being selected for a sample. The number of times it is selected is a random variable t_i with values in the range $t_i = 0, 1, \ldots, k$, and $\Pr(t_i=t) = \binom{k}{t} f^t(1-f)^{k-t}$. If an element is selected t times, it may respond (return the questionnaire) from 0 to t of them. Let $p_{t_i,r}$ be the probability that an element responds r times, given that it has been selected t_i times. Let X_i be a value associated with the i-th element, and consider a statistic

$$x = \sum_{i=1}^{N} u_i X_i.$$

Two methods have been investigated for the choice of the random variable u_i.

a. Method 1

$$\text{Let } u_i = \begin{cases} 2 \text{ if } r \text{ is odd } (1, 3, 5, \ldots.) \\ 0 \text{ if } r \text{ is even } (0, 2, 4, \ldots) \end{cases}$$

Then

$$E\,u_i = E\,E\,(u_i|t_i) = 2\,E\,P_{t_i}$$

$$\left[\frac{t_i-1}{2}\right]$$

where $P_{t_i} = \sum_{j=0} p_{t_i,\,2j+1}$ and the upper limit is the greatest integer less than or equal to $\frac{t_i-1}{2}$.

Then

$$E\,u_i = 2\sum_{t=0}^{k} P_t \binom{k}{t} f^t(1-f)^{k-t} = 2E(P_t) \qquad 1)$$

$$\begin{aligned} \text{Var } u_i &= E\,\text{Var}\,(u_i|t_i) + \text{Var}\,E\,(u_i|t_i) \\ &= E\left[4P_{t_i}(1-P_{t_i})\right] + \text{Var}\,(2P_{t_i}) \\ &= 4\sum_{t=0}^{k} P_t \binom{k}{t} f^t(1-f)^{k-t} \\ &\quad - 4\sum_{t=0}^{k} P_t^{\,2} \binom{k}{t} f^t(1-f)^{k-t} \\ &\quad + 4\left\{\sum_{t=0}^{k} P_t^{\,2} \binom{k}{t} f^t(1-f)^{k-t} \right. \\ &\quad \left. - \left[\sum_{t=0}^{k} P_t \binom{k}{t} f^t(1-f)^{k-t}\right]^2\right\} \\ &= 4E(P_t) - 4E(P_t)^2 = E\,(u_i)[2-E(u_i)] \qquad 2) \end{aligned}$$

b. Method 2

$$\text{Let } u_i = \begin{cases} 2 \text{ if } r = 1 \\ 1 \text{ if } r > 1 \\ 0 \text{ if } r = 0 \end{cases}$$

$$\begin{aligned} \text{Then } Eu_i &= E\,E\,(u_i|t_i) \\ &= E\,(2p_{t_i,1} + 1 - p_{t_i,0} - p_{t_i,1}) \\ &= E\,(1 + p_{t_i,1} - p_{t_i,0}) \\ &= 1 + \sum_{t=0}^{k} (p_{t,1} - p_{t,0}) \binom{k}{t} f^t(1-f)^{k-t} \qquad 3) \end{aligned}$$

$$\begin{aligned} \text{Var } u_i &= E\,\text{Var}\,(u_i|t_i) + \text{Var}\,E\,(u_i|t_i) \\ &= E\,\{E(u^2|t) - [E(u|t)]^2\} + \text{Var}\,(1 + p_{t,1} - p_{t,0}) \\ &= E\,[4p_{t,1} + (1 - p_{t,1} - p_{t,0}) - (1 + p_{t,1} - p_{t,0})^2] \\ &\quad + \text{Var}\,(1 + p_{t,1} - p_{t,0}) \\ &= E\,[p_{t,1} + p_{t,0} - (p_{t,1} - p_{t,0})^2] \\ &\quad + \text{Var}\,(p_{t,1} - p_{t,0}) \\ &= E\,(p_{t,1} + p_{t,0}) - [E\,(p_{t,1} - p_{t,0})]^2 \end{aligned}$$

$$= \sum_{t=o}^{k} (p_{t,1} + p_{t,o}) \binom{k}{t} f^t (1-f)^{k-t} - \left[\sum_{t=o}^{k} (p_{t,1} - p_{t,o}) \binom{k}{t} f^t (1-f)^{k-t}\right]^2 \quad 4)$$

2. Comparison of Methods 1 and 2

We note that

$$Ex = \sum_{i=1}^{N} X_i \, Eu_i = (Eu_i) \sum_{i=1}^{N} X_i$$

$$\text{Var } x = \sum_{i=1}^{N} X_i^2 \text{ Var } (u_i) = \text{Var } (u_i) \; \Sigma X_i^2$$

Hence

$$\begin{aligned} MSE(x) &= \text{Var } x + (Ex - X)^2 \\ &= \text{Var } (u_i) \sum^{N} X_i^2 + (1-Eu_i)^2 \, (\Sigma X_i)^2 \\ &= N\sigma_u^2 \, (\bar{X}^2 + \sigma_X^2) + N^2 \, (1-Eu)^2 \, \bar{X}^2 \\ &= N\sigma_u^2 \, \sigma_X^2 + [\sigma_u^2 + N \, (1-Eu)^2] \, N\bar{X}^2 \end{aligned} \quad 5)$$

This provides a means of comparing the mean square errors of the two methods for hypothetical sets of values $p_{t,r}$.

3. Extension to lists with variable number of duplications.

Let the population be divided in N_1 elements that are on only 1 list, N_2 elements that are on exactly 2 lists, etc. Then, if x_j denotes the estimate for the j-th set

$$x = \sum_{j=1}^{L} x_j = \sum_{j=1}^{L} \sum_{i=1}^{N_j} X_{ji} \quad 6)$$

$$Ex = \Sigma_j \, Ex_j = \sum_{j=1}^{L} E(u|j) \; X_j \quad 7)$$

$$\text{Var } x = \sum_j \text{Var } x_j = \sum_j \text{Var } (u|j) \sum_{i=1}^{N_j} X_{ji}^2 \quad 8)$$

$$\begin{aligned} MSE(x) &= \text{Var } x + (Ex - X)^2 \\ &= \sum_j \text{Var}(u|j) \sum_{i=1}^{N_j} X_{ji}^2 + \{\Sigma_j [E(u|j)-1] X_j\}^2 \\ &= \sum_j \sigma_{uj}^2 \, N_j \, (\sigma_{xj}^2 + \bar{X}_j^2) \\ &\quad + \{\Sigma_j \, [E(u|j)-1] \, N_j \bar{X}_j\}^2 \end{aligned} \quad 9)$$

Formula 9 can be used to compare the efficiency of the two methods.

4. Examples

Two examples are shown in Exhibit I, comparing the two methods for hypothetical populations.

Exhibit I

Data for Example A

Five lists
Sampling fraction = 1/2
Mean per stratum = 1.0
Variance per stratum = 1.0

Stratum: Number of times unit is listed in population	Number of listings	Unduplicated number of units
1	10,900	10,900
2	2,528	1,264
3	477	159
4	108	27
5	90	18
Total	14,103	12,368

Number of times unit is selected (t)	Expected number of units selected	Proportion of times unit responds (r)					
		0	1	2	3	4	5
0	5778.1	1.00					
1	6151.2	0.05	0.95				
2	391.4	0.04	0.50	0.46			
3	32.2	0.03	0.45	0.47	0.05		
4	4.5	0.03	0.40	0.40	0.10	0.07	
5	.6	0.03	0.40	0.35	0.07	0.12	0.03
Total	12,368.0						

Data for Example B

Three lists
Sampling Fraction = 1/2
Mean per stratum = 1.0
Variance per stratum = 1.0

Stratum: Number of times unit is listed in population	Number of listings	Unduplicated number of units
1	139	139
2	84	42
3	12	4
Total	235	185

Number of times unit is selected (t)	Expected number of units selected	Proportion of times unit responds (r)			
		0	1	2	3
0	80.5	1.00			
1	92.0	0.60	0.40		
2	12.0	0.20	0.50	0.30	
3	.5	0.10	0.45	0.40	0.05
Total	185.0				

(continued)

Exhibit 1, continued

Comparisons of Methods 1 and 2
for Examples A and B

Example	Method 1				
	Estimate	Variance	Bias	Mean square error	Relative root mean square error
A	12,116	24,564	-252	88,068	0.024
B	86	259	-99	10,060	0.54

Example	Method 2				
	Estimate	Variance	Bias	Mean square error	Relative root mean square error
A	12,311	23,951	-57	27,200	0.013
B	90	256	-95	9,281	0.52

For both of the examples presented in Exhibit I, method 2 gives a smaller mean square error than method 1.

The relative root mean square error of method 2 for example A is 0.013, while for method 1 it is 0.024. Example B represents an extreme case in many respects. It assumes that:

1. Only about 60 percent of the list corresponds to single units.

2. Only about 40 percent of units selected once responded.

The relative root mean square error for this example is about 0.5 for both methods, because of the extremely biased results obtained. In this example again, method 2 is slightly better than method 1.

D. Reducing the amount of duplication

The unduplication of the various lists for the 1969 Census of Agriculture was done by matching on EI and SS numbers. It was felt that most of the remaining duplication would be removed as a result of the instruction to the respondent who received more than one questionnaire that he should fill out only one, but return them all in the same envelope. This system was adopted because of budget and time considerations and because it was believed that the amount of duplication was small. However, the mailings were made at different times. About 90 percent of the mailing pieces were sent in January 1970, and 10 percent in May and July. Therefore, not all forms were received by respondents at the same time; so respondents were sometimes unable to mail back all forms in the same envelope. In addition, some respondents who should have followed this instruction failed to do so.

An idea of the reasons for duplication can be obtained from Exhibit II-B, which is based on a small sample of "births," which were names added to the mailing list in July 1970. Potential "births" were matched against the original Census list on the basis of SS and EI numbers. While the distributions of duplicates described in Exhibit II-B apply only to the "birth" match, they may provide a general indication of the kinds of duplication problems that are not adequately taken care of by a straight match on identification numbers.

The data in Exhibit II, and results from some other studies of duplication in selected geographic areas, indicate that, if additional characteristics, such as name and address are used in the matching, the amount of duplication may be reduced considerably.

Further unduplication of the mailing lists is necessary for several reasons:

1. To reduce the reporting burden on respondents, and costs to the Census Bureau.

2. To the extent that respondents fail to return duplicate questionnaires, the application of imputation procedures for nonrespondents to these cases may produce a significant upward bias in farm counts and related statistics.

3. Respondents may fill out and return more than one report for the same operation. To the extent that it is not possible to identify these as duplicates, there is an upward bias in all statistics.

In preparation for the 1974 Census of Agriculture, we are investigating various methods of linking records to devise an unduplication procedure which is more effective than the one used for the 1969 Census of Agriculture.

FOOTNOTE

1/ This analysis was suggested by Benjamin J. Tepping of the U.S. Bureau of the Census.

Exhibit II

A. SAMPLE FROM "BIRTH" MAILING LIST: NUMBERS OF DUPLICATED LISTINGS AND OF TOTAL DUPLICATION, BY REGION

Characteristic of the sample	United States	Northeast	North Central	South	West
Total sample size	535	126	134	137	138
Sample listings duplicated one or more times	175	44	44	49	38
Total duplications	267	68	63	79	57

B. PERCENT DISTRIBUTION OF WEIGHTED NUMBER OF DUPLICATES, BY REASON FOR DUPLICATION AND BY REGION

(Based on sample of 267 duplications to a sample of birth listings)

Reason for duplication	United States	Northeast	North Central	South	West
Total	100.0	100.0	100.0	100.0	100.0
Incorrect punching or reporting of SS or EI numbers*	12.3	22.8	9.7	9.2	12.4
SS number on one listing, EI number on other, same name on each	34.1	30.8	28.3	36.6	39.0
Different SS number reported for same name or one listing missing SS number	11.0	4.8	8.0	20.4	5.7
Different EI number reported for same name or one listing missing EI number	10.6	13.5	3.5	16.9	7.6
Other members of partnership or operation	30.6	24.8	46.9	16.9	35.2
No apparent reason	1.5	3.4	3.5	-	-

* SS stands for Social Security and EI stands for Employer Identification.