SOME SAMPLING TECHNIQUES FOR CONTINUING SURVEY OPERATIONS

By: Leslie Kish and Irene Hess
Survey Research Center, University of Michigan

Sample survey methods have been developed largely in terms of individual sampling operations. However, many organizations carry on their survey operations on a rather continuous basis. The special techniques of "rotating" samples devoted to collecting periodically a specific set of data have been treated by Hansen, Hurwitz and Madow, Cochran, Yates, Albert Eckler, and others [1]. We deal here instead with some techniques generally applicable to any repeated use of the same selection frames with similar survey methods. We shall not dwell on the obvious advantages of an adequate staff working as a team experienced in their special problems. Nor shall we discuss the obvious advantages of prorating the costs of expensive sampling materials over many surveys, as was done in the Master Sample of 1945 [3].

We shall present briefly several simple but useful ideas. Probably most of them will appear obvious to statisticians engaged in similar continuing sampling operations. Yet several of these we had to invent -- or more likely reinvent -- for ourselves. This experience and the impressions gained in teaching these methods to others indicate that at least some of these ideas will prove both interesting and useful to some of our readers.

1. The actual sample for each new survey is selected not from the entire population but from a frame of segments and dwellings selected earlier with a predetermined rate from the entire population. The preparation of this frame needs specialized work, done separately from the surveys, preferably during slack periods both in the office and field schedules. We try to do the field work in the months with favorable weather when travelling around the primary areas is easier and somewhat cheaper. In each place we design a frame which can supply us with sampling materials for a period of one to three years. The period is varied and controlled for efficiency considerations; for example, we prepare a smaller frame in places where we expect rapid growth. From this frame we subsample as needed in such a way that the product of the two probabilities of selection -- first, into the frame and second, out of the frame -- equals the probability desired for a particular survey. In the simple situation of uniform monthly surveys (a kind we do not have at the Survey Research Center) the frame can be compiled for a year and then subsampled, one-twelfth for each month.

In addition to the economy and the speed of obtaining samples for specific surveys, subsampling from a frame offers opportunities for greater statistical efficiency through various devices. Of these, we shall describe four briefly. Some of these introduce additional selection stages or phases for obtaining the necessary information for creating good segment boundaries, for assigning measures of size, or for stratification. For example, our interviewers cheaply obtain economic ratings of dwellings, and we have used these for stratified allocation schemes.

a. Instead of sampling area segments directly from a map, we ask the interviewer to create segments out of a "chunk." As viewed on the office work map, the rural chunk is an area which seems to have good boundaries and contains about 30 dwellings; we send the interviewer a chunk sketch and ask her to revise it to reflect the current situation, to indicate the location of each dwelling, and to add any internal features that may be used to divide the chunk into segments of about four dwellings each. In urban areas the chunk may be all or part of a city block, subsampled after a preliminary visit to the block by the interviewer who reports dwelling unit counts and locations, number of floors and apartments in apartment buildings, and related information. This permits us to send the interviewer, for her survey work, segments with boundaries that are familiar to her and with recent measures of size. From the chunk we sample segments without replacement and continue until virtually every segment has been selected for some survey.

b. When we select tracts or blocks from Census listings, we use double sampling; first we select a relatively large sample, stratify it, and then draw from it as needed. It costs scarcely more to select from the Census lists a larger initial sample than a smaller one; thus the cost of drawing the sample can be divided by the number of times it is used. Furthermore, we obtain for the actual sample the gains of stratifying (by geography and by income indicators) the larger initial sample.

c. Currently, we use city directory listings of addresses wherever these are practicable; here, too, we select a larger frame and then subsample repeatedly. Generally, we first select clusters of ten or twenty directory lines, then subdivide these into clusters of about four dwellings for a survey sample. We estimate the probable number of dwellings at each address and group nearby addresses to create clusters of about four dwellings.

   But for a three-year frame of the City of Detroit, we selected clusters of three lines and then subselected single lines for each year's sample; thus obtaining, for each year's sample, individually selected addresses throughout the city.

d. We also use double sampling in the area supplement designed for picking up new growth and missed dwellings. This supplement is particularly necessary to correct selections from city directories. First, we may select an initial sample of blocks eight times as large as we need for a single survey and obtain for these, in the field, rough estimates of size. Then we are free to subselect either one-eighth of the blocks with little growth, or one-eighth of the dwellings from the blocks with much growth.

2. Practitioners of survey sampling know the painful feeling of surprise when they find 20 or even 200 dwellings in a small segment where they expected about four. Area samplers are in a continuing race with home builders. This problem is likely to occur six months or a year after the compilation of the frame of chunks and segments, when the sampler can no longer distribute the surprise building over all subsamples from the frame. For the sake of equal probabilities the sampler might accept all of the surprise dwellings -- and the corresponding increase in variance and cost. At some point, however, he may decide to cut the sample take, thus accepting some bias, but probably with a lower mean square error than the unbiased procedure.

This source of bias could be reduced over the long run of continuing operations by averaging these events in a surprise stratum. The population expansion from a surprise of x_g dwellings (or other elements), from the g-th survey taken with the over-all sampling rate of f_g, is x_g / f_g. The average over G surveys is

$$\left[\sum^{G} w_g x_g / f_g\right] / \sum^{G} w_g ,$$

where w_g is the weight given to the g-th survey. In many situations w_g should be made proportional to f_g; then the average becomes

$$\sum^{G} x_g / \sum^{G} f_g .$$

The sample "take" from the surprise stratum for the last survey should be, then,

$$f_g \sum^{G} x_g / \sum^{G} f_g .$$

The summation G would be over a period (perhaps two years) large enough to provide a "long run" for averaging but not so large as to cause bias by obsolescense. The current survey can be included as the last survey, denoted by G.

We have not used this method since we had only one surprise over the past several years, thanks to an elaborate system of information on growth from the sample counties. This includes interviewers' reports, during surveys, about perceived growth in the visited chunks. But now we intend to lower the criteria for "surprises" and, over several studies, to establish a "surprise stratum." We invite your suggestions and reports of your experiences with this problem.

3. When facing the problem of changing the selection probabilities of a set of sampling units, the sampler may want to use a method that will minimize the number of sampling units that must be changed because changing them is expensive. In particular, we have in mind primary sampling areas, counties or metropolitan areas, each representing an investment of hundreds of dollars in interviewer training and in sampling materials. Nathan Keyfitz has described the problem and a procedure for introducing new population sizes [2].

We may represent the original probabilities used for the selection of sampling units from a stratum as

$$\sum^{J} i_j + \sum^{K} d_k + \sum^{M} s_m = 1 .$$

Similarly, for the same sampling units the new probabilities to which we want to change are

$$\sum^{J} I_j + \sum^{K} D_k + \sum^{M} S_m = 1 .$$

Here i_j and I_j denote the original and the new probabilities of the same sampling unit; there are J + K + M sampling units in the stratum. We use the letters i and I, d and D, s and S to denote, respectively, sampling units with increase, decrease or the same probability from the original to the new measure. These three subsets of sampling units have the relationships

$$I_j > i_j , \qquad D_k < d_k , \qquad S_m = s_m .$$

Also we have $\sum^{M} S_m = \sum^{M} s_m$;

hence, $\sum^{J}(I_j - i_j) = \sum^{K}(d_k - D_k)$.

That is, the sum of the probability increases must equal the sum of the probability decreases.

The procedure for changing probabilities is as follows:

a. If the originally selected sample unit shows either an increase or no change in probability, it remains in the sample with the new probability I_j or S_m .

b. If a sample unit decreases in selection probability, then its probability of remaining is made D_k / d_k and its probability of being dropped is made $1 - \frac{D_k}{d_k}$. If we decide (by resorting to

a table of random numbers) that the unit remains, the compound probability of original selection and remaining is

$$d_k \times \frac{D_k}{d_k} = D_k .$$

c. If a unit is dropped from the sample, we select a replacement from among the increased units with probabilities proportional to the increases, the probability of selection for the j-th unit being

$$\frac{I_j - i_j}{\Sigma(I_j - i_j)} .$$

Thus, the total selection probability for a unit that increased is

$$i_j + \Sigma(d_k - D_k) \times \frac{I_j - i_j}{\Sigma(I_j - i_j)} = I_j .$$

Our methods represent generalizations of the Keyfitz technique in three directions. First, we introduced considerations of statistical efficiency into the problem, knowing that it is neither necessary nor possible to have precise measures of size. It is necessary and sufficient that the sum total of net changes be zero. Within that requirement we can adjust the probabilities of selection to satisfy some criteria of change sufficiently large to be recognized as "important." We noted that for many of the sampling units the change in probabilities was small and unimportant. To these units we reassigned the old probabilities and they became the S units. This procedure reduces the probability of having to switch sampling units; it also eliminates the task of having to revise office records for the sample units with no change in probabilities. In choosing criteria, we tried to balance the increased costs involved in changing primary sampling units (psu's) against the increase in survey variances due to the increased variation in the sizes of sample clusters arising from the small distortions in the probabilities of selection. Of course, we had only crude measures for these criteria, but we believe "anything worth doing at all is worth doing badly." We decided on the following procedure: (a) define important increase as 10 per cent or more and add all such increases over the entire stratum; (b) then add enough decreases from an ordered set of decreases and adjust balance exactly the increases; (c) consider all other sampling units as not having changed. We might have defined a minimum critical decrease, but this we did not consider necessary. Incidentally, rather than merely accepting a specific amount of change from one Census period to another, the rate of change can be projected into the middle of the period of the use of the frame. In other words, the California counties which have increased from 1950 to 1950 will tend to increase through the 1960's; and the sampler making the adjustment in 1960 may take this into account in designing his sample for the '60's.

Second, we also introduced controlled selection into the changes of probabilities. Faced with rather small probabilities of change $(1 - \frac{D_k}{d_k})$ in each of 54 strata (zero in many), instead of drawing independently in each stratum, we cumulated the expected fractions of change from one stratum to another and applied an interval of one, after a random start. Thus, the actual number of changes was controlled within a fraction of the expected number of changes.

To illustrate the application of both the "strict" and the "flexible" assignments of probabilities we display their results for the 54 primary sampling units selected from as many strata in the Survey Research Center's national sample:

	Classification of Sample psu's	Number
	Total primary sampling units	54
A.	Increase with both strict and flexible plans	6
B.	Increase with strict, but remains same with flexible	20
C.	Decrease with strict, but remains same with flexible	12
D.	Decrease with both strict and flexible plans	16

This saves all changes in 32 units, among which 12 were also exposed to being dropped under the "strict" plan. The 6 units with increases were retained in the sample and assigned their new probabilities, with all their records relating to probabilities of selection corrected. Using a controlled selection procedure for dropping psu's resulted in changing 3 of the possible 16. The other 13 required record changes only to convert the old probabilities to the newly adjusted ones.

The illustration in Table 1 of one stratum may clarify some of the details of the flexible procedure. Subset A of psu's includes all with probability increases of 10 per cent or more (see column 4); the new probabilities for these psu's are the same with either plan (columns 3 and 7). Subset B includes psu's with increases of less than 10 per cent (in column 4); these psu's are assigned probability changes of zero in column 6, and the new probabilities in column 7 are identical with the original probabilities (column 2). Next, we must define "decrease" in such a manner that the net change in probabilities will be zero over the entire stratum.

To define "decrease" we order the psu's with respect to the ratio of probabilities (that is, the order of column 4). Beginning with the lowest, proceed up column 5 cumulating the decreases until their absolute values just exceed the sum of the increases, .03640; each decrease (the last five entries in column 5) was adjusted

proportionately so that the decreases of column 6 equal the increases in absolute value. The remaining psu's, subgroup C, are considered to have no change in probabilities.

Third, we want to add that this method may also be used to adjust the probabilities from the original population to some other population. For example, suppose that the psu's were selected with probabilities proportional to numbers of persons and that we now want to sample another population which is distributed somewhat (but not very) differently from the original population -- for example, a population of physicians, or farmers, college students, or Boy Scouts. The same techniques may be used to adjust the probabilities of selection from the original to the newly desired population.

To summarize our modifications of the Keyfitz method: first, we reduce the expected number of necessary changes; second, we reduce the variability of that number; third, we generalize the applicability of the method.

4. The estimators of the variance of survey results are often subject to large variations. This is particularly true for models which use few primary selections (approximate degrees of freedom). For similar items in several surveys, greater precision may be obtained by averaging computations over several succeeding periods. We are conducting investigations of this problem.

5. Another technique available to organizations conducting surveys at intervals with similar methods is the reduction of the effects of nonresponse by simulating an increase in the number of recalls [4]. This technique consists in adding to the addresses of current surveys the nonresponse addresses from similar recent surveys. The replacement addresses should be chosen from surveys using similar respondent units because not-at-homes and refusals among some respondent units may differ from those among others. Refusals may also depend to some extent on survey objectives and questions. The effect of the procedure is about equal to that obtained by doubling the number of recalls, but without the corresponding increase in expense and trouble.

6. The accumulation of evidence on response rates and coverage rates is another advantage of continuing operations. This permits better control of the sample size through the accumulated knowledge of field results. Furthermore, by studying factors associated with varying rates of response, the researcher can learn something about the sources of nonresponse and how to cope with them. Of course, this knowledge does not accumulate automatically but only with planning and labor.

7. One result of continuing operations is the presence of inertia in different parts of the design. For example, in designing some modest size studies, we often find it cheaper and easier to utilize our standard 66 primary sampling areas, or perhaps half of them, than to design and staff six cities, let us say, for the study. This results from having trained interviewers and sampling frames and materials available in our regularly used primary areas. It seems to contradict the usual rule that it is cheaper to use fewer sampling units.

8. Continuing operations also result in a certain conservatism of methods. Some of this is the justifiable result of having available certain good, economic and reliable methods, tested with long experience; and the preceding example can serve as an illustration. However, we suspect that there must also exist many less justifiable types of conservatism, because one naturally thinks first of methods that seem to have worked well enough, that is, without noticeable catastrophes. It is difficult to view new problems with fresh, unbiased eyes. But one should always strive for that fresh point of view and question his familiar methods, trying to separate the seasoned timber from dead wood.

REFERENCES

[1] Hansen, M.H., Hurwitz, W.N., Madow, W.G., Sample Survey Methods and Theory, Vol. I., New York: John Wiley and Sons, Inc., 1953, pp 491-493.

Cochran, W.G., Sampling Techniques, New York: John Wiley and Sons, Inc., 1953, pp 282-290.

Yates, Frank, Sampling Methods for Censuses and Surveys, New York: Hafner Publishing Company, 1953, Second edition, pp 175-182, 233-235, 260-262.

Eckler, Albert R., "Rotation Sampling", The Annals of Mathematical Statistics, Vol. 26, No. 4 (December, 1955), pp 664-685.

[2] Keyfitz, Nathan, "Sampling with Probabilities Proportional to Size", Journal of the American Statistical Association, 46 (March, 1951), pp 105-109.

[3] King, A.J., Jessen, R.J., "The Master Sample of Agriculture", Journal of the American Statistical Association, Vol. 40, No. 229 (March, 1945), pp 38-56.

[4] Kish, Leslie and Hess, Irene, "A 'Replacement' Procedure for Reducing the Bias of Nonresponse", The American Statistician, Vol. 13, No. 4, (October, 1959), pp 17-19.

TABLE 1

COMPARISON OF STRICT WITH FLEXIBLE PLAN FOR CHANGING SELECTION PROBABILITIES OF SAMPLING UNITS IN ONE STRATUM

Item	Original probability	New probability for strict plan	Ratio of probabilities (Col. 3 ÷ Col. 2)	Change in prob. for strict plan (Col. 3-Col. 2)	Flexible plan	
					Change* in probability	New probability (Col. 2 + Col. 6)
1	2	3	4	5	6	7
PSU Classification						
A. Increase with both strict and flexible plans						
1.	.07307	.09000	1.232	+ .01693	+ .01693	.09000
2.	.09407	.11354	1.207	+ .01947	+ .01947	.11354
B. Increase with strict, same with flexible						
1.	.03317	.03636	1.096	+ .00319	.00000	.03317
2.	.04381	.04718	1.077	+ .00337	.00000	.04381
3.	.08915	.09142	1.025	+ .00227	.00000	.08915
4.	.09073	.09083	1.001	+ .00010	.00000	.09073
C. Decrease with strict, same with flexible						
1.	.05500	.05433	.988	- .00067	.00000	.05500
2.	.04258	.04204	.987	- .00054	.00000	.04258
3.	.07297	.07138	.978	- .00159	.00000	.07297
4.	.09719	.09317	.959	- .00402	.00000	.09719
D. Decrease with both strict and flexible plans						
1.	.05059	.04807	.950	- .00252	- .00238	.04821
2.	.02478	.02343	.946	- .00135	- .00128	.02350
3.	.05611	.04894	.872	- .00717	- .00678	.04933
4.	.09263	.07871	.850	- .01392	- .01315	.07948
5.	.08415	.07060	.839	- .01355	- .01281	.07134
Total	1.00000	1.00000	- - -	- - -	- - -	1.00000
Summation by subset						
A.	.16714	.20354	- - -	+ .03640	+ .03640	.20354
B.	.25686	.26579	- - -	+ .00893	.00000	.25686
C.	.26774	.26092	- - -	- .00682	.00000	.26774
D.	.30826	.26975	- - -	- .03851	- .03640	.27186

* Col. 5 entries for the five psu's of class D. were adjusted by a factor of .03640/.03851 = .9452 to obtain the corresponding col. 6 entries.