CHANGING STRATA AND SELECTION PROBABILITIES*

Leslie Kish, The University of Michigan

Summary

Survey samples are often based on primary units selected from initial strata with probabilities proportional to initial measures M_j. But later samples could better be served with new strata and new measures P_j. The difference between the initial and new strata, and measures, may be due to changes in either the population distribution or in survey objectives. Efficiency dictates retaining in the new sample the maximum permitted number of initial selections. A procedure for changing measures within fixed strata is presented first. Then we develop new procedures for retaining the maximum number of selected units within changed strata. Finally, some problems of selecting units with unequal probabilities are explored.

1. Introduction

Unequal selection probabilities are often assigned to sampling units. Especially in the first stage of survey samples, primary sampling units are often selected with probabilities proportional to their size measures, N_j. Then subsampling, within the selected primaries, of final units with probabilities inversely proportional to the N_j, yields a uniform overall selection rate f:

$$\frac{N_j}{c/f} \times \frac{c}{N_j} = f. \qquad (1)$$

The c and f are constants chosen to yield the desired numbers of primary and final units. Typically the selections are made separately in many strata, with one or two (seldom more) primary selections specified from each stratum.

After the initial selection, the primary units may be used for many surveys. But the initial size measures may differ considerably from new measures that would better suit the needs of current surveys. The difference between initial and desired new size measures may be due to differential changes of size among the sampling units, as revealed by the latest Census. But it can also be due to differences in survey populations. For example, the initial measures may have been based on persons, but now one may want to sample a population of physicians, or college students, or farmers, or values of farm products. If measures of size for the new population are available, and if these differ considerably from those used in the initial selection, then one should prefer to use them.

Furthermore, the best strata for the new population may be sufficiently different from the initial strata to justify changing them. The same changes or differences in population distributions, that lead to changing selection probabilities, also will often motivate changing strata.

A brand new selection would mean changing most of the primary units. But continued use of the initial primary units has several advantages. First, they represent important investments in clerical work, and especially in a selected and trained field force. The field force in each county of a national sample may represent a value of many hundreds of dollars. Secondly, using the same primary units decreases considerably the standard errors of comparisons between surveys, (see data by Kish [3]). This problem is acute for panel samples.

To continue using the entire initial set of primary units would restrict the sample to the initial set of selection probabilities and strata. But one can shift to the new probabilities and strata, and yet retain most of the initial primary selections. To make that shift from the initial to a new set of probabilities and strata, with a minimum number of actual changes of primary units, is the purpose of the methods presented.

Section 2 describes a simple method for changing measures of size within fixed strata, modified in section 5 to introduce further considerations of economy. Section 3 presents methods for retaining initial selections from changed strata, and is the main purpose of this paper; these are derived and further discussed in section 4. Finally section 5 treats some related problems in multiple selections with unequal probabilities.

2. Changing Measures Within Fixed Strata

This section presents a procedure for changing from the initial probabilities p_j to new probabilities P_j for a fixed set of sampling units within a stratum. There are D + I units in the stratum. Of these, D receive a decrease in probability, and I receive an increase, which may also be zero. Subscripts

*This research was done at the Survey Research Center, (Institute for Social Research) with the aid of grant USPHS/GM09889-02 from the National Institutes of Health. The author is grateful to Irene Hess, Vinod K. Sethi, and Roe Goodman for their helpful suggestions.

distinguish the two sets of sampling units, so that

$$P_d < p_d \text{ and } P_i \geq p_i, \qquad (2)$$

where $(d = 1, 2, \ldots, D)$ and $(i = D + 1, D + 2, \ldots, D + I)$.

The initial probabilities used for the selection of sampling units from a stratum, and the new probabilities to which we want to switch, both sum to unity in the stratum.

$$\sum^{D} p_d + \sum^{I} p_i = 1 = \sum^{D} P_d + \sum^{I} P_i. \qquad (3)$$

The sum of the probability increases must equal the sum of the probability decreases:

$$\sum^{I} (P_i - p_i) = \sum^{D} (p_d - P_d). \qquad (3')$$

The rules for changing probabilities follow.

a. If the initially selected sampling unit shows an increase (or no change in probability), retain it in the sample, as if selected with the new probability P_i.

b. If the initially selected unit shows a decrease, retain it in the sample with the probability P_d/p_d; that is, assign a probability of $1 - P_d/p_d$ for dropping it. Thus the compound probability of original selection and remaining is made $p_d \times P_d/p_d = P_d$.

c. If a decreasing unit is dropped from the sample, select a replacement among the increasing units with probabilities proportional to their increases. The probability of selection of the i-th unit is $(P_i - p_i)/\sum^{I} (P_i - p_i)$. Thus, the total selection probability of an increasing unit consists of p_i initially, but it is properly increased if any of the D decreasing units had been selected:

$$p_i + \sum (p_d - P_d) \times \frac{P_i - p_i}{\sum^{I} (P_i - p_i)} = P_i. \qquad (4)$$

Often only one sampling unit is selected from each stratum. But the rules are equally valid for two or more fixed numbers of selections from the stratum. This method was first presented by Keyfitz [1]. It is illustrated, amplified and modified in section 5.

3. Changing Stratum Boundaries

When the sampling units of a population are shifted across stratum boundaries, from a set of initial strata to a set of new strata, we face new problems of practical urgency, whose solution motivated our search. A key step consists in separating the procedures of this section from those of section 2. First, stratum changes are solved with section 3 procedures to obtain the preliminary probabilities p_j; then these are changed with section 2 procedures to new final probabilities P_j. Of course, $p_j = P_j$ if the final measures equal the original measures; and section 3 without section 2 suffices for situations when strata are changed but measures are not.

We concentrate here on starting with a single initial selection from each initial stratum and ending with a single final selection in each new stratum, and this represents a limitation. But the number and the sizes of the initial strata may differ from those of the new strata, and this represents an important situation for sample surveys.

For brevity and clarity, strata will denote the new strata. Each of these is composed of several sets, which denote the portions of initial strata it contains. Each set contains either zero or one initial selection. We must assume that the units are sorted into the new strata with procedures that guarantee that the initial selection of some units has no effect on their sorting. This can be guaranteed either with definitions that permit no latitude, or by a performer ignorant of the identity of selected units, or both for extra safety. The guarantee is necessary to ensure that selection probabilities before sorting are equivalent to selection probabilities after sorting into the new strata. Except for this guarantee, there is complete freedom in forming the new strata.

We shall denote with M_j the probabilities that the units received in their own initial strata. These measures were probabilities and summed to 1 in the initial strata, but their sum $\sum M_j = M$ in the new stratum is not generally 1.

This sum varies from stratum to stratum, but for brevity we shall denote it as M, without a subscript for the stratum. The measures M_j must be converted to the preliminary probabilities $p_j = M_j/M$, with $\Sigma\, p_j = 1$. Note that selecting with probabilities proportional to the measures M_j would be equivalent to selecting with probabilities proportional to the preliminary probabilities p_j. Then these could be converted, with section 2 procedures, to the final probabilities P_j. But this would be only a long procedure equivalent to selecting directly with the probabilities P_j. Both procedures disregard our purpose of retaining in the new sample the initial selections.

First Procedure. Suppose now that the stratum consists of several sets, and denote both these sets and their measures as $A + B + C + \ldots = M$. The typical set A has the measure $A = \Sigma_A M_j$, the sum of the measures M_j of the units it contains. First select a set with probabilities proportional to its size, so that set A has the selection probability A/M. If the selected set contains the original selection, accept it as the preliminary selection with $p_j = (A/M) \times (M_j/A) = M_j/M$. If the selected set does not contain the initial selection, select one with probabilities proportional to M_j (or p_j). In either case, compute p_j and convert to P_j, with the section 2 procedures applied to the entire stratum (not to only the selected set).

Understanding this procedure is simple because it depends on only a two-stage selection for the probability $p_j = M_j/M$, and the equivalence (guaranteed) of the possible prior selection with the M_j.

Second Procedure. In the first procedure the sets are selected proportional to their initial probabilities ($\Sigma\, p_j$), and the conversion from p_j to P_j is applied to the entire stratum. The second procedure selects with probabilities proportional to the final probabilities ($\Sigma\, P_j$) in the sets, and the conversion from p_j to P_j is confined within the selected sets. The Census Bureau switched from a 333 area sample based on 1950 data to a 357 area sample based on 1960 data, and described its procedure as follows [5]:

> "For generality, consider a revised stratum in the 357 area design made up of parts from more than one stratum in the 333 area design. In principle, one of these parts can be selected with probability proportionate to 1960 size. If, then, the selected part does not include a PSU previously selected in the 333 area design, a selection can be made from PSU's in that part with probability proportionate to 1960 size. On the other hand, if the selected part contains a sample PSU from the 333 area design, selection from the part can be made by the Keyfitz method [1], maximizing the change of retaining the previously-selected PSU."

Third Procedure. This depends on identifying uniquely each of the initial strata with only one of the new strata. Then section 2 procedures are applied for changing measures. All units in the new stratum that did not belong to the specified initial stratum are treated as newly created, with $p_j = 0$; and some new strata may consist entirely of such new units. On the contrary, units of the initial stratum that did not transfer to the specified new stratum are treated as eliminated from it, with $P_j = 0$. Rules are needed to specify the unique identification of the initial strata with the new strata; they must avoid selection bias and preferably should maximize the retention of selected units.

The first two procedures have the advantages of simplicity, that may facilitate their adaptation to designs with further demands. Two of these deserve mention. First, some designs call for two or more selections from each stratum. Second, in some designs the selections from different strata are not independent, because some form of "controlled selection" or "multiple stratification" is used.

However, those two simple procedures have the disadvantage of not utilizing initial selections present in other sets, when the selected set fails to contain one. This disadvantage is moderate if one set predominates within each of the strata. But the disadvantage increases rapidly with the number of sets of the same general magnitude present in the strata.

Fourth Procedure. This procedure is optimal in the sense that it retains all initial selections, subject to conversions with the section 2 procedures. It retains all initial selections, whenever one is present in any set of the stratum. It requires the assumptions of a single selection from each stratum, and independence of selections between strata. These points are developed, together with the justifications for the following selection rules, in section 4.

a. If no initial selection occurs in any set in the entire stratum, select one unit directly with the new probabilities P_j.

b. If a single initial selection occurs in the entire stratum, accept it as the preliminary selection with $p_j = M_j/M$. Then convert p_j to P_j with section 2 procedures applied in the entire stratum.

c. If two or more sets contain initial selections, select one set with probability y. From the selected set accept the selected unit, and convert from the p_j to P_j with section 2 procedures applied to the entire stratum.

1. When there are only two sets in the stratum, with total measures A and B, select one set with odds inversely proportional to its measure; that is, select set A with probability $y_a = B/(A + B)$, and set B with probability $y_b = A/(A + B)$.

2. When the stratum contains more than two sets, order all sets into a series of dichotomies, with an objective ordering. Whenever both branches of the dichotomy contain selections, choose one with the rules symbolized with:

$$y_a = \frac{B + (a - b)}{A + B} \quad , \qquad (5)$$

where $a = \frac{A}{A'} = \frac{C + D}{C' + D' - C'D'}$

and $b = \frac{B}{B'} = \frac{E + F}{E' + F' - E'F'}$.

To compute the factors A', B', then C', D', and E', F', etc., follow and continue these rules:

If A contains only a single set, A = C and D = 0, then A' = A and a = 1. Similarly for B. If both A and B contain single sets, we have a = b = 1, and the simple case of two sets in the stratum.

If A contains two sets, A = C + D, and C and D contain only a single set each, then C' = C and D' = D and A' = C + D - CD. Similarly for B = E + F.

If C contains two sets C = G + H, then C' = G' + H' - G'H'. If G contains a single set then G' = G; but if G contains two sets, G = T + S, and G' = T' + S' - T'S'. Similarly for H and for D = I + J.

These rules would become cumbersome if carried far. But in practice, rarely will two sets contain selections, when the stratum has many sets.

An objective rule is needed to prevent personal bias in ordering the dichotomies. One reasonable rule is: (a) A set containing more than half of the stratum establishes the first split. (b) Order all others by size; divide them successively into two parts with equal numbers of sets, putting odd numbers into the second half. For example, a stratum composed of a majority set X, and of 9 other sets (denoted with their ranks) would be divided as follows:

$$\{X\}\ \{[(1,2)(3,4)][(5,6)(7,8|9)]\}.$$

4. Justification and Expansion of the Optimum Procedure

To derive the optimal fourth procedure, regard it as the modification of the first procedure, whose justification is obvious. The selection of one of the sets from the stratum with probability proportional to its measure is made to depend on the probability of its containing an initial selection. Based on those probabilities the rules must achieve the equivalent of selecting the set A with probability proportional to A/M.

Let us begin with a stratum containing two sets only, with the measures A + B = M. The presence of only one set would be merely the special case of B = 0. We shall maintain the odds of choosing between the two sets at A/B, because this ratio of the measures is the basis for the selection within the sets. We use the important fact that these measures A and B also represent the probability for each set that it contains the initial selection from the initial stratum. Hence, if we can assume independence between the two probabilities A and B, we have:

$(1 - A)(1 - B)$ = probability that neither set contains a selection (6)

$A(1 - B)$ = probability that set A, but not set B contains a selection

$B(1 - A)$ = probability that set B, but not set A contains a selection

AB = probability that both sets contain a selection.

Our strategy shall be as follows: When neither set contains a selection we choose between the two sets with the desired odds A/B. When selection is present in only one set, we choose it with certainty as desired. But this introduces an imbalance in the odds. When selection is present in both sets, choose set A with odds $y_a/(1 - y_a)$, computed so as to redress the imbalance introduced before.

In other words, the desired odds A/B are achieved over all possibilities by achieving them separately for the case of no selection, and jointly over the case of one or two selections. We need to solve for y_a in the expression that equates the desired odds A/B to the compound probabilities of the occurance of a case and the selecting of the set:

$$\frac{A(1 - B) \times 1 + (1 - A)B \times 0 + AB \times y_a}{A(1 - B) \times 0 + (1 - A)B \times 1 + AB \times (1 - y_a)} = \frac{A}{B} \qquad (7)$$

and $(1 - B) + By_a = (1 - A) + A(1 - y_a)$.

Thus when both sets contain selections, set A should be chosen with probability

$$y_a = \frac{B}{A + B}, \quad \text{or odds} \quad \frac{y_a}{1 - y_a} = \frac{B}{A} \qquad (8)$$

Direct extension of the above method proved too complex even for three sets. But one can arrange a larger number of sets into a series of dichotomies, and then apply in sequence procedures similar to those for two sets. For example, a stratum containing four sets can be divided into group A containing sets C and D, and into group B containing sets E and F. We can choose group A with the odds $A/B = (C + D)/(E + F)$, then select C with odds C/D from $A = C + D$, and finally a unit within C with the probability M_j/C. The overall probability of selecting the unit would be $A/(A + B) \times C/(C + D) \times M_j/C = M_j/M$.

We want an equivalent procedure in which the odds A/B, C/D, and E/F are preserved. A new problem arises in the probability that a group of two sets contains either one or two selections. The probabilities for the two groups are not simply $A = C + D$ and $B = E + F$; they are $A' = C + D - CD$ and $B' = E + F - EF$ respectively. But we can deal more conveniently with the odds A/B, proportional to the sums of measures, than with A'/B'. Hence we must adjust accordingly the probabilities y_a and $(1 - y_a)$ for choosing between the two groups, when both contain selections. We set these so as to attain, jointly for double selections and single selections, the desired odds A/B:

$$\frac{A'(1 - B') + A'B'y_a}{(1 - A')B' + A'B'(1 - y_a)}$$

$$= \frac{Aa'(1 - Bb') + Aa'Bb'y_a}{(1 - Aa')Bb' + Aa'Bb'(1 - y_a)}$$

$$= \frac{A}{B}, \qquad (9)$$

where $A' = Aa'$ and $B' = Bb'$. Then $a'(1 - Bb') + Ba'b'y_a = (1 - Aa')b' + Aa'b'(1 - y_a)$ and $y_a(Aa'b' + Ba'b') = (b' - a') + Ba'b'$.

Thus when both groups A and B contain selection group A should be chosen with probability

$$y_a = [B + (1/a' - 1/b')]/(A + B), \text{ or}$$

$$y_a = \frac{B + (a - b)}{A + B}, \text{ where } a = \frac{A}{A'} \text{ and } b = \frac{B}{B'}. \qquad (10)$$

Thus $(a - b)$ represents an adjustment of the probability of choosing group A over group B, when both contain selections, and when each is composed of two or more sets. Of course, group B would be chosen with $(1 - y_a) = [A + (b - a)]/(A + B)$. Note that a value of $(a - b) > A$ would result in the impossible choice of $y_a > 1$; similarly $(b - a) > B$ would mean $y_a < 0$. But this cannot occur if A and B are both not greater than 3. Since this requirement will be met easily in practice, we desist from pressing further solutions for this problem.

The presence of only three sets in the stratum means that group A contains only one set $A = A'$ and $a = 1$. The presence of only two sets means $a = b = 1$, and (10) becomes (8). When there are only four sets, and $A = C + D$ and $B = E + F$, then

$$A' = C + D - CD \quad \text{and} \quad B' = E + F - EF \qquad (11)$$

Here C and D, and E and F all represent the probabilities for each of the four sets, that it contains one or more selections.

When there are more than four sets in the stratum, the formation of dichotomies can be continued as needed. Any and all of the four groups, C, D, E, and F, can be considered as having been composed of two sets. In general then, instead of (11) we write

$$A' = C' + D' - C'D' \text{ and } B' = E' + F' - E'F' \quad (12)$$

where each of C', D', E', and F' must denote the probability that the group contains one or more selections. For example, if C contains a single set, C' = C; if it contains exactly two sets, C = G + H, then C' = G + H - GH; if both G and H are groups of sets then C' = G' + H' - G'H', where G' and H' are similarly defined.

These procedures for single selections per stratum would become too complex if applied to multiple selections from strata. More direct and general solutions seem to be possible if some strong symmetries are imposed on the strata and on the selections. But generalizations in this direction are not helpful if they impose conditions that are typically absent in practical work.

The simplest example would be the initial selection with simple random sampling of n units from the entire population of N units; that is, all $\binom{N}{n}$ combinations equally probable. These can be sorted into arbitrary strata with procedures that guarantee no effect from the initial selections on the sorting. The n selections can be accepted as equivalent to random selections within the strata. The stratum samples then can be increased or decreased at random to obtain the numbers of random selections needed from each stratum. The symmetries of simple random selections are sufficient, and no more complicated proofs are needed.

If the initial selection was comprised of simple random selections within the initial strata, then they are also random selections within the corresponding sets of the new strata. Treating the sets as substrata within the new strata, some treatment can be evolved to yield a sample approximating a stratified random sample within each new stratum. But for selecting units with unequal probabilities this is likely to become too complicated to be practical.

The problem of selecting two (or more) units with unequal probabilities from a stratum can be simplified by dividing the stratum into two (or more) random substrata of equal size. This can be done with a random sorting of units, as in the first procedure of section 6. Selecting one unit from each substratum reduces the problem to that of a single selection per stratum. The initial strata and the new strata can both be treated this way. We should add that after the single selections in the new strata are secured, other selections can be added either with replacement, or by using one of the procedures of section 6 for selecting without replacement.

Single selections per strata characterize much practical work, and to these the optimal procedures are directly applicable.

The requirement of independence between strata is violated when controlled selection is used. I fear that it would be too difficult to achieve formal treatment of the joint probabilities for the several sets within strata; but this seems most unlikely as a source of genuine bias. However, imposing a controlled selection on the new strata seems difficult to achieve.

5. Modifications and Illustration of Changing Measures within Fixed Strata

a. A simpler procedure than that of section 2 can handle a problem confined chiefly to large growth in a small proportion of the primary units. For example, suppose we have a sample of blocks selected with the initial probabilities p_j, proportional to the initial sizes N_j. Suppose also that for a new sample one is willing to retain the original probability p_j for all blocks, except for those that have at least doubled in size. That is, if the new size $N'_j < 2N_j$, we accept the initial p_j; but for the growth blocks, with $N'_j \geq 2N_j$, we want new probabilities P_j proportional to N'_j.

Generally, place into new growth strata the portion $(N'_j - N_j)$, denoting the size increase of primary units designated as growth units. Then select from these growth units a sample with probabilities proportional to the values $(N'_j - N_j)$. In most situations, the constant of proportionality will be the same (c/f) as for the initial selection, to preserve the uniform overall sampling fraction f; but a different fraction can be introduced for the growth stratum.

Thus the growth units are selected with the sum of two probabilities in the growth strata and in the initial strata:

$$\frac{N'_j - N_j}{c/f} + \frac{N_j}{c/f} \qquad (13)$$

If selected, these primaries should be subsampled with the rates c/N'_j so that the overall selection probability becomes:

$$\left(\frac{N'_j - N_j}{c/f} + \frac{N_j}{c/f}\right) \times \frac{c}{N'_j} = f. \qquad (13')$$

Note that, as desired, the planned subsample is $\underline{c}$ when a unit becomes selected. This can occur both in the initial strata and in the growth strata. The procedure amounts to splitting the growth units into two parts, consisting of an initial size N_j plus a growth size $(N'_j - N_j)$, and subjecting them to separate selections. The ordinary units, that do not qualify as growth units, remain subject to the initial selection rates $N_j/(c/f) \times c/N_j = f$.

b. Into the procedures of section 2 one can introduce further considerations of efficiency, knowing that generally it is neither necessary nor possible to have precise measures of size. It is desirable to have probabilities approximately proportional to size measures. And it is necessary and sufficient for applying the procedures of section 2 that the sum of changes in probabilities be zero in the stratum, (so that $\Sigma P_j = \Sigma p_j = 1$ can be satisfied). Within that requirement we can adjust the selection probabilities to satisfy some criteria of change large enough to be recognized as important, and to deliberately neglect smaller changes.

For many sampling units the changes in probabilities are small enough that they can be deliberately neglected. If these units are numerous, and if neglecting their changes adds little to the overall variance, then one can reduce the number of changes significantly with little sacrifice in the variance. To these units one can reassign the initial probabilities; they constitute a set of units for which $p_i = P_i$ is arbitrarily assigned. This "flexible" procedure reduces the number of units one must switch; it also eliminates the task of revising office records for selected units with no change in probabilities. But this "flexible" procedure requires more computations for balancing the probabilities within the strata than the rigid "strict" plan does. The overall advantage of the flexible plan is important when it confines expensive switching to a small proportion of units. This may require tolerating rather wide limits of differences in actual measures for the units to which $P_i = p_i$ are arbitrarily assigned.

We used the flexible procedure in changing from the 1940 to 1950 Census measures for the 54 counties, representing as many strata in the national sample of the Survey Research Center. In choosing criteria we balance the costs of changing counties against the increase in survey variances due to small distortions in the selection probabilities. These were considered relative to other sources of variations in sample size. We decided on the following procedure: (a) Define an important increase as 10 per cent or more $(P_j/p_j \geq 1.10)$. Compute the sum of the increases in the stratum. (b) Add enough of the largest decreases (smallest values of P_j/p_j) to balance exactly the sum of the increases; that is, $\Sigma (P_i - p_i) - \Sigma (P_d - p_d) = 0$. (c) Consider all other counties as not having changed probabilities, with $P_i = p_i$. The details are in [2].

Two other modifications deserve brief mention. Faced with rather small probabilities of change $(1 - P_d/p_d)$ in each of 16 strata, we did not draw independently within each stratum. Instead we controlled the number of changes by cumulating the probabilities of change from one stratum to another, then applied an interval of one, after a random start. Thus the actual number of changes, which was three, was controlled within a fraction of the expected number of changes.

Second, we need not merely accept the changes from one Census period to another; we can also project them forward into the middle of the period of the use of the Census frame. This may be worth doing for the fastest and steadiest growing counties. For example, the California counties that have shown unusual growth in one decade may also be expected to show similar growth in one decade may also be expected to show similar growth in the next decade. The projection may improve the design of a sample of counties for use during that decade.

6. Some Related Problems

When selection probabilities and strata are changed, the situation may also require changes in the boundaries of sampling units. Some may be split, and others combined. The initial selection probabilities p_j of a unit may be divided among several portions according to specified rules; similarly several of the p_j may be combined into one new unit. The creation of entirely new units is noted with $p_j = 0$, and the elimination of initial units with $P_j = 0$. But we must avoid a detailed treatment of this problem.

Selecting more than one unit with probabilities proportional to unequal measures is difficult to do "without replacement". (It is even more difficult to find simple formulas for computing the lower variances that selection without replacement yields.) This problem has been the subject of several investigations, most recently by Fellegi [6] and Rao [4]. He discusses the first procedure below, but the other four procedures are new, I think. For brevity two selections per stratum are discussed, but the methods can be extended to more selections.

1. Sort the units of the stratum at random into two equal substrata and select one unit from each, with probability proportional to P_j. This can also be accomplished by putting the units in random order and selecting two with a systematic interval.

2. Select from the stratum with equal probability and without replacement 8 units and put them in the same random order in which they were selected. Obtain the sum of their measures, Σ. Now select a random number from 1 to $\Sigma/2$; then add $\Sigma/2$ to it. These two numbers are the two selection numbers from the stratum. The number 8 is suggested merely as a desirable compromise: we want to save labor, but we also want to avoid situations in which double selection is possible, because a unit is larger than $\Sigma/2$.

3. When selecting units with equal probability, reselection of any unit is avoided by substituting some unselected unit. This is simple because of the complete symmetry of equal probabilities, which is generally lacking in units with unequal probabilities. But we can use the presence of partial symmetries, when at least two units appear with the same size, for any size, in the same stratum. If one unit is selected twice, select at random one of the others with the same size. If exactly equal sizes do not exist, generally one can establish strict rules for "best matches" of sizes. When two sizes P_j and P_k are matched, the uniform subsampling rate f may be obtained by subselecting in both with $(P_j + P_k)/2$, whenever both units appear in the sample.

4. First select $K = 6$ units using the desired measures and with replacement, which is generally easy. Then sort these at random into three pairs, but without allowing the same unit twice into a pair. Now select one of the pairs at random. The probability for a unit with $M_j = 0.10$ of appearing twice in two draws is $1/100$; but its appearing more than three times in $K = 6$ draws is reduced to about $1/800$. For $M_j = 0.05$ the reduction is from $1/400$ to about $1/12{,}000$. (If K sampling units are selected with probabilities proportional to measures P_j, subselection of k units with equal probabilities yields a sample of k units with probabilities proportional to the P_j.)

5. An initial selection with equal probabilities would be a special case of the procedures in section 2, with the $p_j = 1/N$ for a single selection, and $p_j = 2/N$ for 2 selections; then convert these to probabilities proportional to the P_j. This procedure reduces but does not eliminate the probability of duplicate selection of the same unit.

References

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

[2] Kish, Leslie and Hess, Irene, "Some Sampling Techniques for Continuing Survey Operations", Proceedings of the Social Statistics Section, American Statistical Association, Washington (1959).

[3] Kish, Leslie, "Variances for Indexes from Complex Samples", Proceedings of the Social Statistics Section, American Statistical Association, (1961), pp. 190-199.

[4] Rao, J. N. K., "On Three Procedures of Unequal Probability Sampling Without Replacement", Journal of the American Statistical Association, 58(March, 1963), pp. 202-215.

[5] U. S. Bureau of the Census, "The Current Population Survey - A Report on Methodology", Technical Paper No. 7, (1963). U. S. Government Printing Office, Washington, p.67.

[6] Fellegi, Ivan P., "Sampling with Varying Probabilities Without Replacement", Journal of the American Statistical Association, 58(March, 1963), pp. 183-201.