ADDITIONAL USES FOR KEYFITZ SELECTION

J. Michael Brick, David R. Morganstein, and Charles L. Wolter, Westat, Inc.
J. Michael Brick, 1650 Research Blvd., Rockville, MD 20850

1. Introduction

In this paper we present two applications of resampling or reselection methods to accomplish different sample design objectives. In the first application, we use a resampling method that maximizes the probability of retaining units previously drawn in a stratified simple random sample. The sampling units may have migrated from one stratum to another, come into existence, or ceased to exist between the time of the initial sample selection and the current one. The method is an extension of the Keyfitz procedures for stratified simple random samples and has the advantage that it is very simple to apply. In the second application, we use Keyfitz procedures to select units for multiple purposes. Units may fall into one or more subgroups and the resampling method is used to draw a sample which has the desired subgroups size and limits the number of distinct units in the sample. Each application is described with a detailed example.

The original idea for retaining units from a previous sample was presented by Keyfitz (1951). The method he devised is applicable to the case in which units remain within a stratum from the time of the original sampling to the next sampling, but their measures of size (probabilities of selection) change. For a one selection per stratum design, Keyfitz gave a method which maximized the probability of retaining the units drawn at the time of the original selection. Retaining existing sampled units in the sample is frequently a very cost effective procedure. Most resampling methods of this type are called Keyfitz methods because of his original contribution.

The methods for the more general case in which the units may also change from one stratum to another were addressed by Perkins (1970) and Kish and Scott (1971). The methods they presented were mainly confined to the case of one selection per stratum, although Kish and Scott did provide methods that are not optimal for some special cases of more than one unit per stratum.

Causey, Cox, and Ernst (1985) formulated the resampling problem as a linear programming problem, in particular as a transportation problem, and gave a general solution for the case of more than one selection per stratum. They noted that the transportation problem can become very large quickly as the number of units selected per stratum increases and may render the solution of the problem in this fashion impractical.

The method presented in the second section handles the specific case of a stratified simple random sample with k units selected per stratum in a very efficient manner. The example given is one in which formulation of the transportation problem is not practical, even though it still is a viable theoretical solution. The method presented is also similar to that given by Kish and Scott (1971) for units selected with equal initial probabilities.

The application of the Keyfitz method to the multiple purpose sample design has a much less well documented record. Kish and Scott (1971) noted the usefulness of resampling methods for this purpose but did not pursue the idea in any detail. The application to this design problem is the focus of the third section.

2. Resampling Methods for Stratified Simple Random Samples

In this section methods for drawing a new, stratified simple random sample to improve its efficiency and to provide an opportunity for updating the frame are presented. The primary method will be illustrated with an example in which the units being selected are post offices. A stratified simple random sample was drawn originally and now a new sample is desired. The problem addressed is how to resample post offices but still retain as many of the originally sampled post offices in the sample. The advantage of retaining the sampled offices is that the data collection and quality control mechanisms are already in place in the offices in the original sample. There is a cost associated with setting up these systems in a new sample post office.

The simplest procedure is to select a completely new sample of post offices independent of the original sample. This is analogous to using the procedures employed in the original sample selection but with the current universe of post offices with their current stratum. The disadvantages of this method are the costs associated with fielding a new sample of post offices (in terms of money and data quality) and the fact that the procedure must be repeated each year the sample is used.

The consequences of a completely new and independent sample selection can be evaluated by computing the expected number of sample offices that would be retained in the new sample and the expected number of new post offices that would be included in the sample. In the example shown later in this section, only 5 percent of the sample offices are retained using this method. The consequences to the operation of the sampling system under this scheme are immense and unnecessary. A different approach can be used to minimize the disruption to the operation of the system.

In the original sampling, a simple random sample was selected within each stratum. Some of the offices have since migrated to different strata, some offices have been eliminated, and some new offices have been established. The objective is to draw a new simple random sample (with a specified sampling fraction) from the post offices in each of the new strata while retaining as many of the original sample post offices as feasible.

As an example of the suggested method, the resampling of post offices for a particular stratum (stratum 5) is described in detail. Westat applied this technique in a study done for the United States Postal Service. Table 1 shows the information needed to determine the conditional selection probabilities of the post offices in the example. The method of computing these probabilities is given below. The reselection probabilities are conditional upon whether or not the post office was drawn in the original sample. If the office was in the original sample, then the conditional probability is the probability of retaining the office in the sample; otherwise, the conditional probability is the probability of selecting the office in its current stratum given that it was not in the original sample from its original stratum. The algorithm produces a sample with the desired unconditional probabilites of selection for the current sample.

The algorithm to determine these conditional probabilities for each post office in the current frame is:

a. Compute the sampling rate or probability of selection for a post office in the original and current samples. Let the original rate be denoted as f_i and the current rate as f_j. Since a simple random sample within stratum is being drawn, the sampling rates are equal to the sample size divided by the universe size for each sample. If a unit is not in the frame for either the original or current sample, then the rate for that office in that year is equal to zero.

b. If the current rate is greater than or equal to the original rate ($f_j \geq f_i$), then assign the post office a conditional probability of selection which depends on whether the office was in the sample originally. If the office was sampled originally, then assign it a probability of selection equal to unity. If it was not in the sample originally, then assign it a conditional probability of selection equal to the quantity (f_j - f_i)/(1 - f_i).

c. If the current sampling rate is less than the original sampling rate ($f_j < f_i$), then assign the post office a conditional probability of selection which depends on whether the office was in the sample originally. If the office was sampled originally, then assign it a conditional probability of selection equal to f_j/f_i. If it was not in the sample originally, then assign it a conditional probability of selection equal to zero.

The conditional probabilities for the post offices in current stratum 5 are shown in Table 1, where the sampling fraction of 0.05 is sought for the current sample. For example, 19 sample post offices in stratum 6 of the original frame migrated into stratum 5 in current frame. These offices have a conditional probability of selection of unity since their current sampling fraction (f_j=0.05) is greater than their original sampling fraction (f_i=0.042). They will all be retained in the sample. The 544 offices in stratum 6 originally that were not in the sample and migrated to stratum 5 currently have a chance of coming into the current sample of (0.05-0.42)/(1-0.42)=0.009. The expected number of new offices that will be selected from this group is 4.7 (544 x 0.009 = 4.7). Note that the numbers shown in the table are rounded, but the calculations are based upon the unrounded numbers.

The expected number of offices that will be retained in the sample and the expected number of new offices that will be drawn in the sample can be computed for stratum 5 by following these procedures over each migration pattern. For this example the expected number of offices that are retained is 43.9 and the expected number of new offices is 16.7. The expected percent retained by this method is 72 percent as compared to the 5 percent retained with an independent reselection.

What remains is to show that the method provides the desired unconditional probabilities of selection for the current sample and that it does this while retaining the maximum number of sample offices. First, we will show that the conditional probabilities produce the desired overall or unconditional rates. This is done separately for the case $f_j \geq f_i$ and $f_j < f_i$.

Case 1: $f_j \geq f_i$

If the post office was sampled at time 1 then it is assigned a conditional probability of being retained equal to unity. If the office was not sampled at time 1, it is assigned a conditional probability equal to (f_j - f_i)/(1 - f_i). The overall probability can then be written as the sum of the products of the probability of selection at time 1 and the conditional probability. For this case the overall probability is:

$$\begin{aligned}\Pr\{\text{in sample}\} &= (f_i)(1) + (1-f_i)(f_j - f_i)/(1 - f_i) \\ &= f_j.\end{aligned}$$

This is exactly the desired probability of selection for post offices in stratum j at time 2.

Case 2: $f_j < f_i$

If the post office was sampled at time 1, then it is assigned a conditional probability of being retained equal to (f_j/f_i). If the office was not sampled at time 1, it is assigned a conditional probability equal to zero. The overall probability can then be written as the sum of the products of the probability of selection at time 1 and the conditional probability. For this case, the overall probability is:

$$\begin{aligned}\Pr\{\text{in sample}\} &= (f_i)(f_j/f_i) + (1-f_i)(0) \\ &= f_j.\end{aligned}$$

Again, we obtain the desired overall probability of the post office being in the sample at time 2.

The resampling method satisfies the objective of updating the selection probabilities of the units so that the overall, unconditional chance of being in the sample after the resampling is equal to the desired sampling fraction. In addition, it does this optimally in the sense that no other conditional probabilities can provide the desired overall sampling fractions and have a greater expected number of

Table 1. Example of resampling algorithm for current stratum 5

Current		Original		Offices in original sample			Offices not in original sample		
Stratum	Probability of selection	Stratum	Probability of selection	Frame	Conditional probability	Expected sample size	Frame	Conditional probability	Expected sample size
5	0.050	3	0.063	0	0.079	0.0	2	0.000	0.0
5	0.050	4	0.217	3	0.231	0.7	15	0.000	0.0
5	0.050	5	0.096	33	0.523	17.2	283	0.000	0.0
5	0.050	6	0.042	19	1.000	19.0	544	0.009	4.7
5	0.050	7	0.018	7	1.000	7.0	232	0.032	7.5
5	0.050	8	0.008	0	1.000	0.0	56	0.042	2.4
5	0.050	9	0.004	0	1.000	0.0	18	0.046	0.8
5	0.050	99	0.000	0	1.000	0.0	26	0.050	1.3
Total						43.9			16.7

retained offices. The proof of this is exactly the same as the proof given by Keyfitz (1951). Suppose unit k falls under Case 1. If it were sampled at time 1, it is retained with certainty which is clearly optimal. If it were not sampled at time 1, the conditional probability must be equal to $(f_j - f_i)/(1 - f_i)$ in order for the overall probability to be f_j. If unit k falls under Case 2, the same logic applies. The units that were not in the sample at time 1 are given zero chance of being sampled. This leaves the largest possible conditional probability of retaining the units that were in the sample at time 1, while still obtaining the desired overall rate. The use of the two cases is necessary because units whose probabilities increase or decrease must be subjected to a chance of being added or dropped from the sample in order to attain the overall rates.

As a final comment on this method, note that the expected sample size for the independent reselection method is 62 offices and for the new method is 60.6 offices. In most cases, the variation about the expected size is not a serious practical problem. If it is necessary to have tight controls on the expected sample size, then some further effort must be employed.

The difference between the observed and desired sample size arises because the observed sample size is subject to rounding error and, more importantly, the conditional probabilities are based upon the number of units that have migrated between strata. The second phenomenon is a random event. The migration pattern of the units is just one of many patterns that could have been obtained. A different original sample of units would have produced a different migration pattern. The resampling algorithm ignores these patterns and the conditional probabilities are assigned solely by the desired sampling fraction assigned for current sample. In the process, the sample size for the current sample is a random variable dependent upon the migration pattern (or equivalently on the sample of units at the first sample).

If it is necessary to tightly control the size of the current sample, then it is possible to modify the resampling algorithm to come closer to the desired size. The modification is described in detail in Brick, Bryant, and Edmonds (1986). It is an iterative procedure in which the sampling rate for current sample is adjusted depending upon the migration pattern that is actually observed.

3. Selecting Units with Multiple Subgroups

In Section 2 we described a common use of resampling methods. This involved the selection of a second, subsequent sample from the same universe where overlap of the second sample with the first is desired and where the measures of size have been updated or units have migrated between strata. Old units may have ceased to exist while new units may have been added to the universe.

In this section, we discuss an application of Keyfitz resampling with a different objective. In this example, units are selected for multiple purposes. Estimates of different subgroups are needed. Each unit contains elements which fall in one or more subgroups and a sample with specified subgroup sizes is required. Examples of such situations might be: estimates of students by school grade where a specific number classes by grade are needed, estimates of university enrollment by program or statistics on individual manufactured product groups. The sampling unit, (the school, university or manufacturer) contains units belonging in one or more subgroups and a sample containing a specified number of subgroup units is desired.

Before describing a solution to this problem involving the Keyfitz method, we describe an alternative approach and discuss its strengths and weaknesses. For the discussion that follows, we will use as an example the Survey of Oral Health in School Children. This survey of public and private school children in elementary and secondary grades was conducted by Westat on two occasions. In the first instance carried out in 1980, a method involving the construction of "pseudo-schools" was used. In the second instance carried out in 1986, a modification of the Keyfitz method was used.

The first stage of sampling involved the selection of 83 PSU's from a national frame. This stage of sampling does not enter into our discussion. The second stage of selection involved the sampling of schools within counties. At a third stage, classes within sampled schools were selected within the PSU so as to provide two classes from each of the 13 grades, kindergarten through twelfth. All students within a class were examined. Survey estimates were performed separately by grade and not aggregated across grades. The sample plan called for the selection of exactly two classes per grade but from different schools in the PSU with no school providing more than four sampled classes.

To prevent burdening schools, the selection of classes was accomplished via stratification by grade ranges. Three grade ranges were used:

(1) K,2,4,6,
(2) 1,3,5, and
(3) 7-12.

In effect, all classes in the PSU were stratified into these three categories and separate samples drawn from each stratum. A school containing all grades K through 6 was randomly assigned to either stratum (1) or (2) to insure that it was not selected from both strata.

The selection of schools had to be performed in such a way that the enrollee sample sizes for the 13 grade levels were approximately equal, about 50 per grade. The characteristics of individual schools was not critical to the objectives of the survey. Schools can be thought of as clusters of grade categories. A sample design which selects schools prior to selecting classrooms from grade categories will be more cost-efficient than a design which randomly selects classrooms from the PSU's universe of classrooms for a particular grade category. In this later case, as many as 26 separate schools may be selected in a PSU. By clustering, we sought to reduce the number of sampled schools per PSU to less than a dozen.

Method 1. The Use of Pseudo-Schools

In the 1980 design, prior to sampling schools in the selected PSU's, the schools were examined for minimum size. Those failing the size criteria were collected together into a larger unit, denoted a "pseudo-school". By size, we refer only to the grade range found in the school and not the numbers of students. Unfortunately, the available data sources only provided us with a total school enrollment figure and a grade range, not a count of students by grade. Individual grade enrollment was assumed to be uniformly distributed across the grades found in the school. Using this estimate of enrollment, "pseudo-schools" were then selected using a probability proportional to their combined enrollment.

Our objective in the formation of a "pseudo-school" was to create a sampling unit capable of providing a minimum sample. For example, in grade-range stratum (1), a school had to provide at least four classes, one in each grade K, 2, 4, and 6. If it did not, it would be collected with another school in the PSU to make a unit meeting this criteria. After the aggregation process, any "pseudo-school" sampled from the grade-range (1) stratum can be assured of providing the required four classes.

Considering the need for two classes from each of the grades, six pseudo-schools had to be selected, two from each of the three grade-range strata. Since each pseudo-school contained one or more schools, the actual number of schools selected per PSU was at least six and often as much as twelve.

The advantage of using the "pseudo-school" approach is that it limits the number of schools which must be contacted within the constraint of controlling the number of sample classes per school. The disadvantage in this approach is the amount of clerical work needed to create the "pseudo-schools". In most PSU's, all schools must be examined for their grade range offering and where necessary combined with another one or more other schools.

A more serious drawback to the pseudo-school approach involved substitution for non-cooperation or erroneous grade-range data. Occasionally, a school administrator would refuse participation in the study. As it turned out, the grade-range data was occasionally in error and even the "pseudo-school" could not supply the requisite sample of classes. Either condition resulted in messy substitutions and schedule delays since several schools had to be contacted to obtain the cooperation for the entire replacement "pseudo-school". To enroll a new "pseudo-school" cooperation had to be obtained from as many as four principals and their respective school boards.

Method 2. Use of the Keyfitz Procedure

Because of the drawbacks of the "pseudo-school" approach, an alternative selection procedure was used to draw the 1986 sample of schools. This procedure involved a successive application of the Keyfitz procedure to grade categories. We still faced the same objectives: controlling the number of schools to be contacted in each PSU; achieving two sample classes per grade, no more than one class per grade level per school, a minimum of two classes selected per school, and an average between three or four classes selected per school. This approach allowed us to determine the probability of a particular school providing a class in a given grade category.

In the selection of schools, the Keyfitz procedure was used to maximize the overlap of schools selected in the grade categories applicable to the particular grade-range strata defined earlier. A selection of one school/grade took place within each stratum. A second selection for a different grade category followed, using the Keyfitz procedure to maximize the chances of retaining the school selected at the first step. This procedure was repeated for all grade categories associated with the particular grade-range stratum.

A demonstration of the appropriateness of this method follows from a recursive argument applied to the Keyfitz method for selecting one unit from a stratum using unequal probabilities as discussed in Kish and Scott (1971). Consider the selection probability for a unit's inclusion in the first subgroup, a particular grade, as f_1 and in the second subgroup, a different grade, as f_2. The Keyfitz method insures that the selection of a unit to provide a class for the second subgroup will yield a probability selection with the desired probability f_2. Repeat the same argument for the choice of a school to provide a third class using the probabilities f_2 and f_3 to denote the original and current probabilities, and so forth for subsequent choices.

An Example

To implement the Keyfitz selection procedure, a measure of size was associated with each school within each grade category present in the stratum. Since there were no enrollment counts available by grade for schools, the same assumption used in the 1980 study was made, namely that the enrollment for a school was uniformly distributed across the grades found in the school's grade range. The enrollment for a grade of a particular school was then the school's total enrollment divided by the number of grades found in the school's grade range.

To describe the implementation of the Keyfitz procedure, consider the schools in a grade stratum with four grades, denoted G1, G2, G3, G4. The probability of selection for the ith school within the jth grade category is denoted b_{ij}.

Table 2. School probabilities by grade category

School ID	G1	G2	G3	G4
001	b_{11}	b_{12}	b_{13}	b_{14}
002	b_{21}	b_{22}	b_{23}	b_{24}
N	b_{N1}	b_{N2}	b_{N3}	b_{N4}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1.000	1.000	1.000	1.000	

The steps for selecting the sample of schools using the Keyfitz procedure successively across the grade categories are :

1. In G1, one school is sampled using the probabilities b_{11}, b_{21}, b_{31}, ... b_{N1}. This initial subgroup selection corresponds to a "prior" selection. Suppose the t1th school was the chosen school. From the t1th school, one classroom in grade G1 is selected.

2. The Keyfitz procedure is now used to select a school in the second subgroup, G2, where the t1th school is treated as the initial selection. If the t1th school was retained (either because its probability increased or its probability decreased, it was retained when the random choice was made), one classroom in grade G2 would be selected from this school. If the t1th school probability decreased and it was rejected, then a new school is selected from the collection of schools in G2 whose G2 selection probabilities are greater than or equal to their G1 selection probabilities. This newly selected school would furnish one classroom in the G2 category.

3. The school selected in step 2 now becomes the initial selection for purposes of repeating the Keyfitz procedure. The same decision rules used in step 2 are applied to select a school in G3, but the selection probability b_{t2} is used in the application of these rules. Here t2 denotes the subscript identifying the sample school in G2.

4. If the t3rd school is chosen in G3 to provide one classroom for this grade category, then it serves as an initial selection when the Keyfitz decision rules are executed to yield a selection in G4. The t3rd school is either retained or rejected. If retained, this school provides a classroom from the G4 category. If rejected, one school is selected from the collection

of schools in G4 whose b_{j4} values are greater than or equal to their corresponding b_{j3} values.

The expected number of sample schools can be reduced if the selection probabilities of each school increase from grade category to grade category. This can be accomplished by ranking the grade categories by the frequency of schools in the stratum possessing each of the grade categories applicable to the stratum, with the grade category having the highest frequency of schools coming first in the ranking. For example, suppose a stratum contains five schools having the following grade ranges, and that the grade categories from which schools are to be selected are K, 2, 4, and 6.

1.	K	1	2	3					
2.	K	1	2	3	4	5	6		
3.	K	1	2	3	4	5	6		
4.							6	7	8
5.						5	6		

Assume that the frequency distribution for the grades K, 2, 4, and 6 is:

Grade category	Number of schools having the grade
K	3
2	3
4	2
6	4

The ordering of the grades should be 6, K, 2, and 4, that is, the first selection should be a sixth grade class, the next a Kindergarten, and so forth.

This school sampling procedure should lead to the school-related design specifications described earlier. Since each schools appears in only one elementary group and one secondary group, a sample school can never have more than one class per grade. Since each elementary grade appears in two of the four elementary groups and each secondary grade appears in both secondary groups, there will be two classes for each grade. Because the Keyfitz procedure is being used to maximize the retention of a sample school for two adjacent grade categories, in most cases a sample school will provide at least two classrooms. By dividing the elementary grades into two groups (K, 2, 4, 6 and 1, 3, 5) an elementary school will provide a maximum of four classes. It is possible for sample schools spanning primary and secondary grades to be selected in several grade categories but only if it is selected in both an elementary group and a secondary group.

As a summary comparison, we note that the use of pseudo-schools resulted in an average of between eight and nine schools being selected per PSU. Using the design criteria, the absolute minimum number would be six schools, two schools from each of the three grade-range strata. Using the Keyfitz method to increase the chances of retaining a school once selected, the cluster size increased only slightly to an average of about 10 schools per PSU. The cost of this modest increase in sample was a substantial reduction in the initial cost to select the sample and an improved ability to provide replacements thereby permitting the fieldwork to maintain a tight time table during the school year.

4. Summary

In this paper we have used resampling procedures to accomplish two different objectives. In the first application we have presented a typical Keyfitz type of resampling algorithm that is appropriate for stratified simple random sampling. The algorithm is very easy to use and it gives the optimal probabilities for keeping already selected units in the sample. A derivative of this method which controls the actual sample size by stratum is referred to in the discussion.

The second application uses the normal Keyfitz method for drawing one unit per stratum with updated and unequal selection probabilities. The significant factor in this example is the use of the method for drawing a sample that serves several different purposes. The Keyfitz method is used to increase the chance that a unit sampled for one subgroup can be retained for another subgroup.

REFERENCES

Brick, J.M., Bryant, E.C., and Edmonds, H.J. (1986), "Evaluation of the First-Stage Sampling and Estimation Procedures for the Domestic Revenue, Pieces, and Weight System," Final Report for the Statistical Design Branch, U.S. Postal Service.

Causey, B.D., Cox, L.H., and Ernst, L.R. (1985), "Applications of Transportation Theory to Statistical Problems," *Journal of the American Statistical Association*, 80, 903-909.

Keyfitz, Nathan (1951), "Sampling with Probabilities Proportionate to Size: Adjustment for Changes in Probabilities," *Journal of the American Statistical Association*, 46, 105-109.

Kish, Leslie, and Scott, A. (1971), "Retaining Units After Changing Strata and Probabilities," *Journal of the American Statistical Association*, 66, 461-470.

Perkins, Walter M. (1970), "1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilites Within 1970 NSR Strata," memorandum to Joseph Waksberg, U.S. Bureau of Census.