MULTISTAGE KEYFITZ UPDATING

Douglas J. Drummond, Research Triangle Institute

1. Overview

Sample designs calling for the selection of one primary sampling unit (PSU) per first-stage stratum using probabilities proportional-to-size (pps) abound in the statistical literature. In some cases, this initial selection of units is used to support multiple surveys over a period of several years. In 1951, Nathan Keyfitz [1] introduced a procedure for updating the sample PSUs in such designs to realize selection probabilities proportional to more current data, while maximizing the number of common PSUs in the two samples. This procedure is equally applicable to designs required to support two study objectives, each having an optimal size measure, by separate pps samples having maximum overlap. In this paper, the Keyfitz procedure will be employed at multiple stages of the sample design. To facilitate discussion of this extension, a general representation of multiphase pps selection procedures will first be presented. This representation will then be illustrated using the Keyfitz methodology as well as that of independent selections. Finally, an overview is provided of a recent study carried out by the Institute in which the methodology expounded upon in this paper was employed in the design of a large national program evaluation.

2. General Representation of an Optimal Multi-Phase PPS Updating Strategy

We would like to select one unit per stratum using size measure x for the initial sample and one unit per stratum using size measure X for the second (updated) sample so as to maximize the expected overlap in the two samples. Notationally, PSU i in stratum $\ell(i=1,2,\ldots,N(\ell)$; $\ell=1,2,\ldots,L)$ will have inclusion probabilities given by

$$\left.\begin{aligned} p_i(\ell) &= \frac{x_i(\ell)}{x_+(\ell)} \\ \text{and} \qquad & \\ P_i(\ell) &= \frac{X_i(\ell)}{X_+(\ell)} \end{aligned}\right\} \quad \ell=1,2,\ \ldots,\ L$$

for the initial and updated samples, respectively. A three-phase sampling procedure will be used independently in each stratum in order to potentially condition the second sample by what was drawn in the initial sample for that stratum. Specifically,

Phase 1: Select one unit in stratum ℓ with probabilities proportional to x.

Phase 2: Decide whether unit from Phase I will be retained in second sample.

$$\left[\Pr\left\{\begin{array}{l|l} \text{PSU i in} & \text{PSU i in} \\ \text{stratum } \ell & \text{stratum } \ell \\ \text{retained at} & \text{was selected} \\ \text{Phase II} & \text{at Phase I} \end{array}\right\} = R_i(\ell)\right]$$

Phase 3: If rejection at Phase II, select the second sample stratum ℓ member from PSU frame according to

$$\Pr\left\{\begin{array}{c|c} \text{PSU j of} & \text{PSU i in} \\ \text{stratum } \ell & \text{stratum } \ell \\ \text{selected into} & \text{rejected at} \\ \text{second sample} & \text{Phase II} \end{array}\right\} = S_{j\cdot i}(\ell).$$

Then, under this three-phase sampling procedure,

$$\Pr\left\{\begin{array}{l} \text{PSU i of stratum } \ell \\ \text{is selected into} \\ \text{the initial sample} \end{array}\right\} = p_i(\ell)$$

and

$$\Pr\left\{\begin{array}{l} \text{PSU i of stratum } \ell \\ \text{is selected into} \\ \text{the second sample} \end{array}\right\} = p_i(\ell)\ R_i(\ell) + \Pi_{\bar{R}}(\ell)\ S_i(\ell)\quad ,$$

where

$$\Pi_{\bar{R}}(\ell) = \Pr\{\text{rejection in stratum } \ell \text{ at Phase II}\}$$

$$= 1 - \sum_{i=1}^{N(\ell)} p_i(\ell)\ R_i(\ell)$$

and

$$S_i(\ell) = \Pr\left\{\begin{array}{l|l} \text{PSU i of stratum } \ell & \text{rejection in} \\ \text{selected into the} & \text{stratum } \ell \text{ at} \\ \text{second sample} & \text{Phase II} \end{array}\right\}$$

$$= \left\{\sum_{j=1}^{N(\ell)} S_{i\cdot j}\ p_j(\ell)\ (1-R_j(\ell))\right\} \div \Pi_{\bar{R}}(\ell)\ .$$

Moreover, the number of distinct PSUs in the two samples from stratum ℓ, $n(\ell)$, has expectation given by

$$E\{n(\ell)\} = 1 + C_+(\ell)\ ,$$

where

$$C_+(\ell) = \Pr\left\{\begin{array}{l} \text{distinct units in the two} \\ \text{samples from stratum } \ell \end{array}\right\}$$

$$= \sum_{i=1}^{N(\ell)} C_i(\ell)\ p_i(\ell)$$

and

$$C_i(\ell) = \Pr\left\{\begin{array}{l|l} \text{distinct units in} & \text{PSU i of stratum} \\ \text{the two samples} & \ell \text{ selected for} \\ \text{from stratum } \ell & \text{initial sample} \end{array}\right\}$$

$$= (1-R_i(\ell))\ (1-S_{i\cdot i}(\ell))\quad .$$

To retain the required pps structure for the second sample we must require admissibility, i.e.,

$$p_i(\ell)\, R_i(\ell) + S_i(\ell)\, \Pi_{\bar{R}}(\ell) = P_i(\ell) \qquad \forall i \quad .$$

Moreover, for the sequential updating procedure to be optimal, we would like $C_+(\ell)$ to be minimized ($\ell=1,2, \ldots, L$). Finally, in anticipation of applying updating procedures at multiple stages of the design (when applicable) it can be shown that

$$\Pr\left\{\text{PSU } i \text{ of stratum } \ell \text{ is in both samples}\right\}$$

$$= p_i(\ell)\, R_i(\ell) + p_i(\ell)\, (1-R_i(\ell))\, S_{i\cdot i}(\ell) \ ,$$

$$\Pr\left\{\begin{matrix}\text{PSU } i \text{ of stratum } \ell \text{ is only in the initial} \\ \text{sample}\end{matrix}\right\}$$

$$= p_i(\ell)\, (1-R_i(\ell))\, (1-S_{i\cdot i}(\ell)) \ ,$$

and

$$\Pr\left\{\begin{matrix}\text{PSU } i \text{ of stratum } \ell \text{ is only in the second} \\ \text{sample}\end{matrix}\right\}$$

$$= \sum_{\substack{j\neq i\\ 1}}^{N(\ell)} S_{i\cdot j}(\ell)\, (1-R_j(\ell))\, p_j(\ell) \ .$$

3. Comparison of Two Admissible Multi-Phase PPS Updating Strategies

Two updating strategies will be considered: (a) Independent updating; and (b) Keyfitz updating.

Underlying parameters and properties are given in Table 1. Notice that under the Keyfitz procedure, if a unit is rejected at Phase II it cannot be re-selected at Phase III. It is readily shown that this is a necessary condition in order for any candidate multi-phase pps updating strategy to be optimal (i.e., minimize $C_+(\ell)$, $\ell=1,2,\ldots,L$). Furthermore, employment of independent as opposed to Keyfitz updating in stratum ℓ leads to an excess in expected sample size, $d(\ell)$, given by

$$d(\ell) = \sum_{i=1}^{N(\ell)} \min\{p_i(\ell),P_i(\ell)\}\ [1-\max\{p_i(\ell),P_i(\ell)\}]$$

$$\geq 0 \quad .$$

That is, among these two admissible pps updating procedures, Keyfitz can do no worse than independent selections with respect to the expected number of distinct units selected from each stratum in the two samples. That Keyfitz updating is indeed optimal in this regard is easily shown (e.g., [2]). Moreover, this optimality is actually at the unit level (i.e., change probabilities for every unit of every strata are minimized among the class of admissible pps updating procedures). Finally, Table 2 evaluates the probability of selected simple events that will be of interest in discussing the potential for using the Keyfitz procedure at multiple stages of the sample design.

4. Extension of Keyfitz Procedure to Multiple Design Stages

In some multistage designs utilizing an admissible pps updating strategy at the first-stage of sample selection, one second-stage unit (SSU) is selected with probability proportional-to-size in each secondary stratum for each PSU selected at the first-stage. In cases where the updating procedure was successful in realizing a common PSU for the two first-stage samples, it may again be advantageous to attempt to maximize the overlap in the SSU sample members (assuming common secondary stratification).[1] In considering such an extension, there are sixteen possible sequences (nine of which are feasible under strict multistage sampling) for SSU j of second-stage stratum m in PSU i of first-stage stratum ℓ to enter the underlying samples--Table 3 provides the details. The probability of each of these simple events of interest can then be expressed using information contained in Table 2--Table 4 provides the details for Keyfitz updating.

Notation used in Table 4 extends that of Section 2 of this paper in the obvious fashion. Specifically,

$p_{j\cdot i}(\ell,m)$ = probability that SSU j in secondary stratum m of PSU i in first-stage stratum ℓ is selected into the initial sample conditional on PSU i having been selected into the initial sample

and

$P_{j\cdot i}(\ell,m)$ = probability that SSU j in secondary stratum m of PSU i in first-stage stratum ℓ is selected into the update sample conditional on PSU i having been selected into the update sample

Compound events of interest and their probabilities (after possible algebraic simplication) are given in Table 5. Clearly, each of the underlying samples achieves the desired pps structure. Moreover, since the Keyfitz updating procedure is optimal at each stage of application, the two-stage Keyfitz updating scheme proposed is also guaranteed of maximizing the expected overlap in the second-stage sample units. Finally, the methodology under consideration can easily be extended to more than two stages and/or to allowing the use of Keyfitz updating at one stage and independent updating at the next (and vice versa) provided one unit per stratum is selected at all component stages.

Table 1: Parameters Associated with Independent and Keyfitz Updating[a/]

Parameter of Interest	Type of Parameter	Updating Procedure	
		Independent Samples	Keyfitz
$R_i(\ell)$	Defining	0	$\min\left(1, \frac{P_i(\ell)}{p_i(\ell)}\right)$
$\Pi_{\bar{R}}(\ell)$	Induced	1	$\sum_{i=1}^{N(\ell)} \max\{0, p_i(\ell) - P_i(\ell)\}$
$S_{i \cdot j}(\ell)$	Defining	$P_i(\ell)$	$\max\left\{0, \frac{P_i(\ell) - p_i(\ell)}{\Pi_{\bar{R}}(\ell)}\right\}$
Pr {PSU i of stratum ℓ selected into the update sample} [b/]	Induced	$P_i(\ell)$	$P_i(\ell)$
$C_i(\ell)$	Induced	$1 - P_i(\ell)$	$1 - R_i(\ell)$
$E\{n(\ell)\}$	Induced	$2 - \sum_{i=1}^{N(\ell)} p_i(\ell)\, P_i(\ell)$	$1 + \Pi_{\bar{R}}(\ell)$

[a/] Given the defining parameters $\{R_i(\ell),\ S_{i \cdot j}(\ell)) : i,j = 1,2,\ldots,N(\ell)\}$ the induced parameters are easily derived using the notation and findings of the previous section.

[b/] Candidate procedures are required to have a pps initial sample with respect to size measure x.

Table 2: Probability of Selected Events Under Independent and Keyfitz Updating[a/]

Event	Probability of Event Assuming	
	Independent Updating	Keyfitz Updating
PSU i of Stratum ℓ is in both samples	$p_i(\ell)\ P_i(\ell)$	$\min\ \{p_i(\ell),\ P_i(\ell)\}$
PSU i of stratum ℓ is only in the initial sample	$p_i(\ell)\ [1-P_i(\ell)]$	$\max\ \{0,\ p_i(\ell)-P_i(\ell)\}$
PSU i of stratum ℓ is only in the second sample	$[1-p_i(\ell)]\ P_i(\ell)$	$\max\ \{0,\ P_i(\ell)-p_i(\ell)\}$
PSU i of stratum ℓ not selected in either sample	$[1-p_i(\ell)][1-P_i(\ell)]$	$1-\max\ \{p_i(\ell),\ P_i(\ell)\}$

[a/] Probability of events are obtained by direct substitution into the formulae of the previous section.

Table 3: Listing of all Simple Events Involving SSU j of PSU i Under Strict Multistage Sampling

Simple Event #	PSU i Selected		SSU j of PSU i Selected		Feasibility[a]
	Initial Sample	Update Sample	Initial Sample	Update Sample	
1	Y	Y	Y	Y	Y
2	Y	Y	Y	N	Y
3	Y	Y	N	Y	Y
4	Y	Y	N	N	Y
5	Y	N	Y	Y	N
6	Y	N	Y	N	Y
7	Y	N	N	Y	N
8	Y	N	N	N	Y
9	N	Y	Y	Y	N
10	N	Y	Y	N	N
11	N	Y	N	Y	Y
12	N	Y	N	N	Y
13	N	N	Y	Y	N
14	N	N	Y	N	N
15	N	N	N	Y	N
16	N	N	N	N	Y

[a] Feasibility refers to requirement that for a strict multistage design, in order for SSU j of PSU i to support a given sample, PSU i must also support that sample.

Table 4: Probability of Simple Events Under A Two-Stage Keyfitz Procedure

Feasible Simple Event Number[a]	Probability
1	$\min\{p_i(\ell),\ P_i(\ell)\}\ \min\{p_{j.i}(\ell,m),\ P_{j.i}(\ell,m)\}$
2	$\min\{p_i(\ell),\ P_i(\ell)\}\ \max\{0,p_{j.i}(\ell,m)-P_{j.i}(\ell,m)\}$
3	$\min\{p_i(\ell),\ P_i(\ell)\}\ \max\{0,P_{j.i}(\ell,m)-p_{j.i}(\ell,m)\}$
4	$\min\{p_i(\ell),\ P_i(\ell)\}\ [1-\max\{p_{j.i}(\ell,m),P_{j.i}(\ell,m)\}]$
6	$\max\{0,p_i(\ell)-P_i(\ell)\}\ p_{j.i}(\ell,m)$
8	$\max\{0,p_i(\ell)-P_i(\ell)\}\ \{1-p_{j.i}(\ell,m)\}$
11	$\max\{0,P_i(\ell)-p_i(\ell)\}\ P_{j.i}(\ell,m)$
12	$\max\{0,P_i(\ell)-p_i(\ell)\}\ \{1-P_{j.i}(\ell,m)\}$
16	$1-\max\{p_i(\ell),\ P_i(\ell)\}$

[a] See Table 3 for description of simple events.

5. Application of Multistage Keyfitz Updating

The Research Triangle Institute is currently completing a study of the Title I Migrant Education Program for the Office of Program Evaluation within the newly formed Department of Education. The study is comprised of five main components involving an impact study, funding validation, estimation of funding undercoverage, description of students receiving services, and description of sites providing services, respectively. These components are supported by a consolidated multistage design involving multiple samples, which in possible combination, provide coverage of each underlying target population of interest. No attempt will be made to provide a detailed discussion of this design at this time (see [3] for such an account). In stead, the intent of the present section will be to merely indicate areas of the design where methodology developed in this paper either was employed or at least viewed as a possible alternative to what was implemented. Specifically, Keyfitz updating was used at the second- and third-stages of sample selection with respect to the Impact and Validation components of the study. Secondly, consideration was given to pooling these resulting sample SSUs or TSUs to support the remaining components. Sampling weights in such an application would then be based on a direct extension of Table 5. Thirdly, overlap estimators were proposed for the Funding Undercoverage Component but were rejected in light of the detrimental effect on precision of the positive correlations assumed to exist between the eventual dependent samples of classrooms within each school. These possibilities are exponded upon in [2], where, in addition, the merits of multiple objective Keyfitz updating are contrasted with the properties of a single composite size measure design.

6. Conclusions

Methodology for imposing Keyfitz updating at multiple stages of sample selection was developed in this paper. This extension was found to be tractable (in terms of the frame information required for computing sampling weights), and flexible (in terms of enabling Keyfitz-induced dependencies between samples to be properly accounted for under pooling of such samples and/or the employment of overlap estimators). The additional survey economies available under multistage Keyfitz updating further extend the utility of this important and proven sample selection mechanism.

FOOTNOTE

[1] If a common PSU is not realized, independent second-stage samples are selected in the two PSUs using the size measure under which the component PSU was selected into the first-stage sample.

REFERENCES

[1] Keyfitz, Nathan, *Sampling With Probabilities Proportional to Size: Adjustment for Changes in the Probabilities*. JASA (1951), 46, 105-109, (1951).

[2] Drummond, Douglas J., *The Keyfitz Sample Selection Procedure Revisited*. Expanded version of contributed paper presented at the 1980 Annual ASA Meetings.

[3] Cameron, B.F., Drummond, D.J. et.al. *OMB Clearance Package for Study of the ESEA Title I Migrant Program* (9 November, 1977). Prepared for Office of Evaluaton and Dissemination, Department of Education.

Table 5. Compound Events of Interest Under a Two-Stage Keyfitz Procedure

Compound Event	Component Feasible Simple Event Numbers	Probability of Compound Event
SSU j in secondary stratum m of PSU i in primary stratum ℓ is in initial sample	1,2,6	$p_i(\ell)p_{j.i}(\ell,m)$
SSU j in secondary stratum m of PSU i in primary stratum ℓ is in updated sample	1,3,11	$P_i(\ell)P_{j.i}(\ell,m)$
SSU j in secondary stratum m of PSU i in primary stratum ℓ is only in the initial sample	2,6	$\min\{p_i(\ell),P_i(\ell)\}\ \max\{0,p_{j.i}(\ell,m)-P_{j.i}(\ell,m)\}$ $+\ \max\{0,p_i(\ell)-P_i(\ell)\}\ p_{j.i}(\ell,m)$
SSU j in secondary stratum m of PSU i in primary stratum ℓ is only in the updated sample	3,11	$\min\{p_i(\ell),P_i(\ell)\}\ \max\{0,P_{j.i}(\ell,m)-p_{j.i}(\ell,m)\}$ $+\ \max\{0,P_i(\ell)-p_i(\ell)\}\ P_{j.i}(\ell,m)$
SSU j in secondary stratum m of PSU i in primary stratum ℓ is in both samples.	1	$\min\{p_i(\ell),P_i(\ell)\}\ \min\{p_{j.i}(\ell,m),P_{j.i}(\ell,m)\}$