Survey Methodology, December 1989
Vol. 15, No. 2, pp. 253–260
Statistics Canada

# Updating Size Measures in a PPSWOR Design

ALAN SUNTER[1]

ABSTRACT

It is sometimes required that a PPSWOR sample of first stage units (psu's) in a multistage population survey design be updated to take account of new size measures that have become available for the whole population of such units. However, because of a considerable investment in within-psu mapping, segmentation, listing, enumerator recruitment, *etc.*, we would like to retain the same sample psu's if possible, consistent with the requirement that selection probabilities may now be regarded as being proportional to the new size measures. The method described in this article differs from methods already described in the literature in that it is valid for any sample size and does not require enumeration of all possible samples. Further, it does not require that the old and the new sampling methods be the same and hence it provides a convenient way not only of updating size measures but also of switching to a new sampling method.

KEY WORDS: PPSWOR; Sample updating; PPS sequential sampling.

## 1. INTRODUCTION

It is sometimes required that a PPSWOR sample of first stage units (psu's) in a multistage population survey design be updated to take account of new size measures that have become available for the whole population of such units. This occurs, for example, when the psu's are census enumeration areas (or collections of census enumeration areas) and a new census has made new population/housing counts available or when, because of observed uneven growth in EA populations in an intercensal period, it is decided to do an interim update of size measures in a sampling stratum. However, because of a considerable investment in within-psu mapping, segmentation, listing, enumerator recruitment, *etc.*, we would like to retain the same sample psu's if possible, consistent with the requirement that selection probabilities, originally proportional to the old size measures, may now be regarded as being proportional to the new ones. A comprehensive treatment of the problem for $n = 1$ is given by Kish and Scott (1971) and is itself a generalization of a method given earlier by Keyfitz (1951). They point out that their method may be extended without difficulty to with replacement sampling (PPSWR) for $n > 1$. Their method may also be used (Drew, Choudhry, and Gray 1978; Platek and Singh 1978) for $n > 1$ when the PPSWOR procedure used is that due to Rao, Hartley and Cochran (1962), since this method involves the formation of $n$ random groups and subsequent selection of a single psu from each group. It breaks down however if we wish, as indeed we probably would, to form new random groups according to the new size measures. Fellegi (1966) provides two methods applicable to a PPSWOR sample of $n = 2$ drawn by the Fellegi (1963) procedure.

The method given in this paper is similar to the second Fellegi method and, when applied to the examples in the Fellegi paper, gives very similar results. Unlike that method, however, it does not require the enumeration of all possible samples and hence is a feasible procedure for any value of $n$ and $N$. Although it is formally applicable to any PPSWOR method for which it is feasible to calculate the selection probability of any sample selected it has its highest utility for PPSWOR methods in which all, or nearly all, $n$-tuple subsets are possible samples with

[1] Alan Sunter, President, A.B. Sunter Research Design & Analysis Inc., 63 Fifth Av., Ottawa, Canada, K1S 2M3.

probabilities approximately proportional to the product of their unit probabilities. The method of this type, used for purposes of illustration, is the author's pps sequential method (Sunter 1986, 1989).

## 2. REPLACEMENT PROCEDURE THEORY

We wish to reselect a PPSWOR sample, originally selected with probabilities $\{\pi_{11}, \pi_{12}, \ldots, \pi_{1n}\}$ proportional to original size measures $\{z_{11}, z_{12}, \ldots, z_{1n}\}$ under a new set of probabilities $\{\pi_{21}, \pi_{22}, \ldots, \pi_{2n}\}$ proportional to new size measures $\{z_{21}, z_{22}, \ldots, z_{2n}\}$. However, we want to do this in such a way that we have a high probability of retaining the original sample.

We assume that for any particular $n$-tuple $S$, including of course $S'$, the original sample actually selected, it is possible to calculate both $P_1(S)$, its selection probability under the original scheme, and $P_2(S)$, its selection probability under a new scheme. For many samples in many schemes (*e.g.* pps systematic) one or both of these probabilities may be zero although, obviously, $P_1(S')$ cannot be zero.

The procedure is as follows:

Step 1: (a) Calculate $P_1(S')$, $P_2(S')$.

(b) If $P_2(S') \geq P_1(S')$ then retain the sample.

(c) If $P_2(S') < P_1(S')$ retain the sample with probability $P_2(S')/P_1(S')$. If rejected proceed to Step 2.

Step 2: (a) If the original sample was not retained then draw a new sample, $S_1$ say, with probability $P_2(S_1)$. If $P_2(S_1) < P_1(S_1)$ then reject the sample, otherwise retain with probability $1 - P_1(S_1)/P_2(S_1)$. If rejected proceed to Step 2(b).

(b) If the Step 2(a) sample was not retained then draw a new sample, $S_2$ say, and proceed as for Step 2(a).

(c), (d), ... Repeat the Step 2(a), 2(b), ... procedure until a sample is retained.

The sample eventually retained by this process has the required probability structure for both unit probabilities and unit pair joint probabilities. In other words, it may be regarded as having been drawn under the new scheme. In particular, since it has the same joint probability stucture, it has the same sampling variance.

Let $P^*$ denote the probability that the process does not terminate at Step 1, $P^{**}$ the conditional probability that it does not terminate at Step 2(a) given that it did not terminate at Step 1. Obviously $P^{**}$ is then also the conditional probability that the process does not terminate at any subsequent step given that it did not terminate at any step preceding that step. We now have

$$P^* = \sum_{i:P_2(S_i) < P_1(S_i)} (1 - P_2(S_i)/P_1(S_i))P_1(S_i)$$

$$= \sum_{i:P_2(S_i) < P_1(S_i)} (P_1(S_i) - P_2(S_i)) \qquad (1)$$

where $i$ now indexes the $n$-tuple subsets of the $N$ population units, and

$$P^{**} = 1 - \sum_{i:P_1(S_i) < P_2(S_i)} (1 - P_1(S_i)/P_2(S_i))P_2(S_i)$$

$$= 1 - \sum_{i:P_1(S_i) < P_2(S_i)} (P_2(S_i) - P_1(S_i)) \qquad (2)$$

while, since $\sum_i P_1(S_i) = \sum_i P_2(S_i) = 1$, it is easy to see that the summation terms on the right of (1) and (2) respectively must be equal and we have $P^* = 1 - P^{**}$.

Denoting ultimate selection probability by $P'$ we now have, by design:

For $i:P_2(S_i) < P_1(S_i)$

$$\begin{aligned} P'(S_i) &= P_1(S_i)\ (P_2(S_i)/P_1(S_i)) \\ &= P_2(S_i), \text{ as required.} \end{aligned}$$

For $i:P_2(S_i) \geq P_1(S_i)$

$$\begin{aligned} P'(S_i) &= P_1(S_i) + P^*(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*P^{**}(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*(P^{**})^2(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*(P^{**})^3(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + \ldots \\ &= P_1(S_i) + P^*(P_2(S_i) - P_1(S_i))/(1 - P^{**}) \\ &= P_2(S_i) \end{aligned}$$

as required.

Finally, we observe that the expected number of Step 2 "trials", given that the original sample was not retained at Step 1, is given by the binomial waiting time distribution as $1/(1 - P^{**}) = 1/P^*$.

## 3. APPLICATION AND EXAMPLES

The new scheme need not be the same (even apart from the change in unit probabilities) as the old one. We could switch, for example, from a sample originally drawn under pps systematic sampling to one drawn under the author's (Sunter 1986, 1989) pps sequential scheme or even from PPSWR (pps with replacement) to a PPSWOR scheme. In the latter case, of course, an original sample with multiple inclusions of a single psu has zero probability of selection in the new PPSWOR scheme. The procedure may still be used, it may be noted, even if we have included new psu's in the stratum but are retaining the same sample size.

The procedure probably has its highest practical utility, as measured by its probability of retaining the same sample, when both the old and the new schemes are such that all, or nearly all, samples are possible and their probabilities are approximately proportional to the product of their unit selection probabilities. Under these circumstances, and provided that the changes in size measures are not extreme, $P_1(S_i)$ and $P_2(S_i)$ tend to have about the same values so that the probability of retaining the same sample will be relatively high. A practical PPSWOR method with the required properties is the author's, referred to above. Since we will use this method in the examples of the next section, we now describe it. There are two variants, in both of which we have to find a suitable ordering of the population and accumulate the size measures (which we assume to be scaled to sum to 1), in reverse order (so to speak), to give:

$$Z_i = \sum_{i}^{N} z_j; \; i = 1, 2, \ldots, N.$$

**Variant 1:** Order the population in any way such that

(a) $nz_i \leq Z_i; \; i = 1, 2, ..N - n$

(b) $(n - i)z_i < Z_i; \; i = n, n + 1, \ldots, N - 1.$

Then select units until exactly $n$ have been selected according to:

$$P(U_i \mid n_i) = \begin{cases} 1 \text{ if } n_i = N - i + 1 \\ n_i z_i / Z_i \text{ otherwise} \end{cases}$$

where $n_i$ is the number of sample units still required to be selected when we arrive at the $i$-th population unit.

It is always possible to satisfy the ordering requirements (a) and (b). For example ordering by increasing size obviously satisfies both as does ordering by decreasing size down to the point (if any) at which (b) fails and then by increasing size. The latter ordering has some advantage in that it tends to minimize the slight (and, for practical purposes, negligible) deviation from strict pps for the last $n$ units (see Sunter 1986). Variant 2 avoids these deviations altogether by taking advantage of the fact that if it occurs that there are $n_i + 1$ units remaining in the population for any $i$, then it is usually possible to simply discard one of these units with appropriate probability and retain the others.

**Variant 2:** Order the population in any way such that

(a) $nz_i \leq Z_i; \; i = 1, 2, ..N - n - 1$

(b) $(n - i)z_j < Z_i; \; j \geq i \geq N - n.$

Then

(i) select according to $P(U_i \mid n_i) = nz_i/Z_i$ until either $n_i = 0$ or $n_i = N - i$, then

(ii) if $n_i > 0$ discard one of the remaining units, say that indexed $j$, with probability $1 - n_i z_j / Z_i$ and select the others.

An algorithm for finding an ordering satisfying the requirements for Variant 2 is given in Sunter(1986) and is incorporated in the program used for the simulations of the next section. In both variants $\pi_{ij}$ maybe calculated according to

$$\pi_{ij} = n(n - 1)z_i z_j \tau_{ij}$$

where $i < j$ (in the indexing of the ordering actually used) and

$$\tau_1 = 1/Z_2$$

$$\tau_i = (1/Z_i + 1)(1 - z_1/Z_2) \ \ldots \ (1 - z_{i-1}/Z_i).$$

These expressions are exact for $i < j \leq N - n + 1$, and provide a very close approximation otherwise. They are easily calculated and give the method the advantage, unique among practical procedures for PPSWOR with $n > 2$, of the availability of variance estimation with negligible bias.

Pascal-like pseudocode for a routine that selects a sample according to Variant 1, at the same time calculating its probability and the value of $\tau_i$ for each selected unit, is given in an Appendix. It is easily extended to Variant 2 or modified to the calculation of $P$(S) for an already selected sample.

### 3.1 Example 1

To illustrate these procedures we take first an example with $n = 2$ and $N = 4$, small enough for sample enumeration and manual calculation, where it will be seen that, in order to obtain the "new" size measures, we simply inverted the order of the original assignment. The Variant 2 ordering algorithm mentioned above gives (4,1,2,3) for the first set of size measures and (1,4,3,2) for the second. There are six possible samples, listed in column (1) of Table 2, whose probabilities under the Variant 2 algorithm are easily calculated, with results shown in columns (2) and (3). Column (4) gives the probability of retaining this sample at Step 1, given that it was the original selection. Column (5) gives the conditional probability of retention at any subsequent Step 2, given that no sample was retained at a preceding step.

It may be verified that the overall probability of retention of the same sample, given by the sum of the products of the values in columns (2) and (4), is 0.5465. This value may be compared with the overall probability of retention of the same sample when the new sample is selected independently, given by $\sum_i P_1(S_i)P_2(S_i) = 0.1168$. Thus even in this rather extreme example, we have considerably increased the likelihood of retaining the same sample.

**Table 1**

Selection Probabilities

| PSU | $z_{1i}$ | $z_{2i}$ |
|---|---|---|
| 1 | 0.15 | 0.35 |
| 2 | 0.20 | 0.30 |
| 3 | 0.30 | 0.20 |
| 4 | 0.35 | 0.15 |

Table 2

| (1)<br>Sample | (2)<br>$P_1(S)$ | (3)<br>$P_2(S)$ | (4)<br>$P_{2\|1}(S)$ | (5)<br>$P_{2\|2}(S)$ |
|---|---|---|---|---|
| 1,2 | 0.0231 | 0.3231 | 1.0 | 0.9286 |
| 1,3 | 0.1154 | 0.2154 | 1.0 | 0.4643 |
| 1,4 | 0.1615 | 0.1615 | 1.0 | 0 |
| 2,3 | 0.1615 | 0.1615 | 1.0 | 0 |
| 2,4 | 0.2154 | 0.1154 | 0.5357 | 0 |
| 3,4 | 0.3231 | 0.0231 | 0.0715 | 0 |

### 3.2 Example 2

In a more realistic set of examples we now take $n = 4$ for a population of 100 psu's with "original" size measures independently assigned from the uniform or rectangular distribution $R(1,3)$. "New" size measures are assigned in a number of ways, described below. For these examples it is no longer feasible to enumerate all possible samples or to perform the sample selection and sample probability calculations manually. However, writing a computer program to do the latter and to apply the reselection procedure was a straightforward task. The program was used to perform 200 iterations, for each example, of selection of a sample using Sunter's Variant 2 with probabilities proportional to the first set of size measures with subsequent application of the procedures described above for reselection of a sample with probabilities proportional to the second set of size meaures. The program, running on an XT-compatible operating at 7.16 MHz, generated and sorted the populations of size measures and performed 200 iterations of the sample selection and reselection in about three minutes.

Case 1, in which we have assigned new size measures from the same distribution independently of their original values, may be seen as a "worst practical case" scenario. Case 2, in which 10% of the psu's have doubled in size with the rest remaining unchanged, is an approximation of a "scattered development" scenario. Case 3 illustrates the random perturbation of size measures by an amount rectangularly distributed over an interval equal to the original size measure. From Table 3 it may be seen that with probabiliities ranging from 0.67 in the "worst case" scenario to 0.81 in the "scattered development" scenario, we retain the original sample. For those cases in which the original sample is rejected the average number of Step 2 trials required to select a new sample agreed closely with the predicted value of $1/P^*$.

Table 3

200 Iterations of a Size Measure Update Procedure, $n = 4, M = 100$; Original Size Measures from $R(1,3)$

| Case | Source of $\pi_{2i}$ | Step 1<br>Retentions | Average<br>Step 2<br>Trials | Estimated<br>$P^*$ |
|---|---|---|---|---|
| 1 | $z_{2i} \approx R(1,3)$ | 134 | 2.98 | 0.33 |
| 2 | $z_{2i} = 2^*z_{1i}$ for 10% of psu's | 153 | 5.53 | 0.19 |
| 3 | $z_{21} = R(z_{1i}/2,3z_{1i}/2)$ | 154 | 4.17 | 0.25 |

## ACKNOWLEDGEMENTS

The author would like to thank the Editor and two referees for helpful comments and the correction of errors in the original presentation.

## APPENDIX

### Pseudocode for Variant 1 of PPS Sequential Sampling

It is assumed here that the population of size measures has already been given a suitable ordering, say by the algorithm given in Sunter (1986) and that its index, $i$, in this ordering identifies the unit. Size measures, scaled to sum to 1, are stored in an array $z$[1 .. PopSize] with their cumulative values (accumulated from PopSize down to 1) stored in an array $Z$[1 .. PopSize]. The meaning of the variables will be clear from the names that they are given. The results are to be stored in an array Sample [1 .. SamSize,1 .. 3] in which the elements are population index $i$, unit probability $\pi_i$, and $\tau_i$ respectively. "Random" is a function that returns a random number uniformly distributed on the the interval (0,1). The indentations in the code written below are intended to facilitate the visual pairing of the begin/end's that delineate a compound statement.

```
{Variables initialization}
   i = 1; SamProb = 1; NumRem = SamSize; Gamma = 1/Z[2];
   {Sampling routine}
   while NumRem > 0 do
begin
   if i > 1 and i < PopSize then
      Gamma = Gamma*(1 - z[i - 1]/Z[i])*Z[i]/Z[i + 1];
   if i = PopSize - NumRem + 1 or Random < = Numrem*z[i]/Z[i]
   then
   begin
      if i < > PopSize - NumRem + 1 then
         SamProb = SamProb*NumRem*z[i]/Z[i];
      NumRem = NumRem - 1;
      Sample[SamSize - NumRem,1] = i;
      Sample[SamSize - NumRem,2] = SamSize*z[i];
      Sample[SamSize - NumRem,3] = Gamma;
   end else SamProb = SamProb*(1 - NumRem*z[i]/Z[i]);
   i = i + 1;
end.
```

## REFERENCES

DREW, J.D., CHOUDHRY, G.H., and GRAY, G.B. (1978). Some methods for updating sample survey frames and their effects on estimation. *Survey Methodology*, 4, 225-263.

FELLEGI, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

FELLEGI, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, 434-442.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size. *Journal of the American Statistical Association*, 58, 183-201.

KISH, L., and SCOTT, A., (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, 24, 482-490.

PLATEK, R., and SINGH, M.P. (1978). A strategy for updating continuous surveys. *Metrika*, 25, 1-7.

SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.

SUNTER, A.B. (1989). PPS Sampling in multistage designs: does it matter which method? Manuscript submitted to *Journal of Official Statistics*.