

## COORDINATION OF PPS SAMPLES OVER TIME

Esbjörn Ohlsson, Stockholm University  
Mathematical Statistics, Stockholm University, S-106 91 Stockholm, Sweden  
esbj@matematik.su.se

### ABSTRACT

Probability proportional to size (PPS) sampling finds application in business surveys both for the first stage of a multi-stage design, and for direct sampling of businesses from a list frame. In both cases there is a need to update samples from time to time, while retaining as many units as possible from the old sample. In the second type of application there is also a need for negative coordination of surveys to get an even distribution of response burden. Further requirements on the sampling procedures are simplicity in application and in estimation of variance. We present various permanent random number techniques that meet these requirements and compare them to a few other methods, and present a simulation study on expected overlap.

**Key Words:** Positive Coordination, Negative Coordination, Overlap Control, Permanent Random Numbers

### 1. THE PROBLEM

Applications of PPS sampling in business surveys can be split into two main categories: (a) sampling from area frames with a multi-stage design and (b) sampling ultimate units directly from a list frame. The classical situation for (a) is a multi-stage sample where primary sampling units are drawn with probability proportional to some measure of the unit's size (PPS). Typically, due to extensive stratification, just a few units are selected in each stratum.

As an example we consider a master sample of stretches of road, maintained by the Swedish National Road Administration. Even though the first stage sample size is 84, extensive stratification results in stratum sample sizes  $n=1$ , throughout. In each sampled unit, investments are made, e.g. on equipment to measure traffic flow, and this, of course, is a strong argument for retaining the sample over the years. Another argument is that the main interest in samples from this frame is in estimates of change over time, for which retaining the same units improves precision.

The sizes of the units are based on a rotating census that gives estimates of traffic mileage per year for one third of the units each year. Since these sizes changes substantially over the years, there is a regular need to update the samples, if not every third year so at least each decade. Else, there would finally be a great loss in efficiency in the estimates from surveys using the master sample. Furthermore, at irregular intervals there are changes in the classification of roads that underlies the stratification: e.g., roads may change from county to national or even European highway status. Also, there are some new roads (births) and roads that are no longer in the population (deaths).

We conclude that in this survey, as in most repetitive surveys, there is a need for updating the sample to account for new sizes, stratum/classification changes plus births and deaths, while retaining as many units as possible from the old sample.

In this example, the within strata sample sizes were  $n=1$ , as is the case with several major surveys such as the US Consumer Expenditure Survey and the US Current Population Survey. This might explain why this case has received considerable attention in the literature see e.g., Keyfitz (1951), Kish and Scott (1971) and Causey, Cox and Ernst (1985). Several newer references can be found in the overview by Ernst (1999). In the present paper we will put a special emphasis on the case  $n=1$ , but we will also consider PPS samples of other.

The second type of application of PPS sampling is the case where we have a list frame with ultimate sampling units, often in the form of a business register. The most common design in this case seems to be a stratified simple random sample (STSI). Sample coordination with this kind of design was discussed extensively at the 1993 ICES meeting, see Ohlsson (1995) and Srinath and Carpenter (1995). It is also a topic at the ICES II meeting, see McKenzie and Gross (1999) and Royce (1999). In these cases, there is both stratification by industry and by size. The PPS alternative means that the stratification by industry is kept while stratification by size is replaced by PPS sampling. One advantage of this kind of design is that the size measure is more extensively exploited.

One example of a PPS design of this second kind is the sample of outlets for the Swedish CPI. In an investigation of the sample of retail traders for the Swedish Consumer Price Index, Ålenius (1990) considered the sample of  $n=41$  department stores out of  $N=259$  units in that industrial stratum, with prices of two different commodities as target variables. With a standard size stratification, using four strata, STSI gave twice the variance of PPS sampling. Even when the STSI design was refined to use 41 strata and a ratio estimator was used (treating size as an auxiliary variable), STSI gave around 25 percent larger variance than PPS. For the NASS Crops Survey, Bailey & Kott (1997) found PPS sampling to be more efficient than STSI for most crops.

The conclusion is that there are business surveys for which a PPS design is preferable to STSI from an efficiency point of view. Furthermore, if the PPS sampling procedure is simple to implement, a PPS design is simpler to administrate, with much fewer strata to construct, allocate and maintain. Even if we do not claim PPS to be generally preferable to STSI, it should be considered a strong candidate for business survey designs. A necessary condition for this is that the PPS procedure involved is simple to use in practice. Simple and efficient PPS procedures are the main topic of this paper.

## 2. SPECIFICATION FOR PPS COORDINATION

There are an immense number of PPS procedures available in the literature, see Brewer & Hanif (1983). Most of these are not proper for sample coordination, though. In fact, we believe that lack of simple, efficient PPS procedures that can produce coordinated samples may be a reason for the extensive use of STSI instead of PPS sampling. We now specify in more detail our requirements on a simple and efficient PPS sampling procedure with capability of handling sample coordination.

The (stratum) population is  $U=1,2,\dots,N$ . In the frame (which is supposed to be a list even in case a) there is a non-negative auxiliary variable  $p_1, p_2, \dots, p_N$ . In applications,  $p_i$  is usually a measure of the size of unit  $i$ . We assume that the  $p_i$ 's have been normed so that  $\sum p_i = 1$ , within each stratum.

Särndal, Swensson and Wretman (1992, p. 90) give a list on desirable properties of a PPS procedure. Applied to the situation with two samples that are to be coordinated, their first two properties give us the following three requirements:

- (i) Relative simplicity in application.
- (ii) For the first sample  $\pi_i = \Pr(i \in s) = np_i, i \in U$ .
- (iii) For the second sample  $\pi'_i = \Pr(i \in s') = n'p'_i, i \in U'$ .

Here  $\pi_i = \Pr(i \in s)$  denotes actual inclusion probability of unit  $i$  in the first sample  $s$ . All quantities relating to the second sample  $s'$  will be equipped with a prime, as in  $p'_i$ .

Särndal et al. add three conditions that enable variance estimation with the Sen-Yates-Grundy estimator. In a later article, Särndal (1996) argues for the use of procedures that allow simple, single-sum variance estimation. We agree with this point of view and get

- (iv) Availability of a variance estimator, preferably expressed as a single sum. (Not relevant for  $n=1$ .)

Finally, we add three conditions that are particular for the problem of overlap control. Note that the expected number of units in common to two samples is

$$\sum_i \Pr(i \in s, i \in s') \quad (1)$$

- (v) Possibility of positive sample coordination of two or more samples with different size measures  $p$ , different strata and different  $n$ , preferably with maximization of the expected sample overlap in (1).

- (vi) Possibility of negative sample coordination, of two or more samples with different size measures  $p$ , different strata and different  $n$ , preferably with minimization of the expected sample overlap in (1).
- (vii) On each occasion, all strata are sampled independent of each other. For the second sample this means that  $\Pr(i \in s', j \in s') = \Pr(i \in s') \Pr(j \in s')$  whenever  $i$  and  $j$  are in *different* strata.

As noted by Ernst (1999), condition (vii) is not satisfied by most overlap procedures that allow for different stratification. This condition is important for obtaining unbiased variance estimates and, above all, for the possibility to apply the sample overlap procedure repeatedly.

Särndal et al. (1992, p. 90) remark that it is not easy to devise a (fixed-size) procedure having the properties desirable for PPS sampling, even at a single occasion. It is thus futile to hope for a procedure for sampling on two occasions, with overlap control, that fulfils all our requirements (i)-(vii). Instead, we have to look for procedures that are reasonably good at (i)-(vii). We start out with a procedure that is not fixed-size, i.e. it gives a random sample size.

### 3. POISSON SAMPLING AND THE IDEA OF PRN

In Poisson sampling, each unit is given an independent, uniformly distributed random number  $X_i$  on the interval (0,1). Unit  $i$  is included in the sample  $s$  if  $X_i \leq np_i$ . Poisson sampling can be used for sample coordination by saving the  $X_i$  as *permanent random numbers* (PRN). This idea is due to Brewer, Early and Joyce (1972) and means that when a second sample is to be drawn, we use the same random numbers as in the first, but we update the sizes  $p$ , the stratification, and the sample size  $n$ . Note that the quantities  $p'_i$  and  $n'$  relates to the stratum where  $i$  is located in the second design, which may or may not be the same as in the first design. (For simplicity in notation, we refrain from using stratum sub-indexes.)

A virtue of Poisson sampling is that it is very simple to apply, i.e. requirement (i) met. Further, it is readily seen that this procedure is strictly PPS so that (ii) and (iii) are fulfilled. A single-sum variance estimator (iv) is available, see Brewer and Hanif (1983, p. 83). The probability of including unit  $i$  in two samples drawn with PRN Poisson sampling is obviously

$$\Pr(i \in s, i \in s') = \min(np_i, n'p'_i) \quad (2)$$

This is of course the largest possible probability, yielding a strict maximum in (1). Negative coordination can be achieved by shifting the PRN an amount  $c$  to the right before the selection of the second sample, giving new random numbers  $X_i^* = X_i + c$ . If  $m$  samples are to be negatively coordinated, the choice  $c = 1/m$  should give a small sample overlap. In particular, if the target inclusion probabilities  $np_i$  are less than  $1/m$  for all units  $i$  in all  $m$  designs, the expected overlap is 0. In the case of  $m=2$ , an alternative is to use antithetic random numbers,  $X_i^* = 1 - X_i$  which gives minimum expected sample overlap for any target probabilities.

We conclude that both (v) and (vi) are satisfied. Finally, all units are sampled independently and in particular (vii) is met.

It may appear as if we have found the optimal procedure for PPS sample coordination. However, Poisson sampling has the drawback of giving a random sample size (with expected value  $n$ ). This has two implications: The first is that there is a risk for  $n=0$  in some stratum. Since the random sample size is approximately Poisson distributed, the probability of this to happen when we have  $H$  strata, all with sample size  $n$ , is  $1 - (1 - e^{-n})^H$ . In order for this quantity to be negligible we must avoid small  $n$ , where the magnitude of "small" of course depends on  $H$ . The conclusion is that Poisson sampling can not be used in all situations, and in particular not in those with  $n=1$  or  $2$ . Even when the probability of some zero sample size is negligible, the randomness in sample size may seriously disturb the intended sample allocation over strata, with a loss in efficiency of the estimates.

The second, less serious, drawback of the random sample size is that it should be clear that inference should be made conditional on the sample size actually obtained. Conditional on the actual sample size, the probabilities of inclusion are no longer exactly PPS and they are in fact very hard to compute, see Aires (1999).

#### 4. FIXED SIZE PROCEDURES

We now look at fixed-size alternatives to Poisson sampling, starting with the case  $n=1$ . In this section we restrict the attention to PRN procedures. Collocated sampling, by Brewer et al (1972), cannot be used to coordinate samples with different stratification, and will therefore not be considered here.

##### 4.1 The case $n=1$

A sampling design is a probability distribution on all possible sets of samples. With this definition, PPS sampling with fixed size  $n=1$  is unique. The traditional sampling procedure for realizing a PPS size one sample uses just one random number. For the application of PRN sample coordination, we need a procedure that uses individual random numbers for all units. Such a procedure is *Exponential sampling*, presented in Ohlsson (1996). Starting with a set of PRN,  $\{X_i; i=1,2,\dots,N\}$ , we compute the transformed random numbers  $\xi_i = -\log(1 - X_i) / p_i$ , which are exponentially distributed with mean  $1/p_i$ . The unit with the smallest  $\xi_i$  is selected for the sample. By a well-known result from probability theory, the probability of selecting unit  $i$  is  $p_i$ , as required.

Coordination of samples is achieved by using PRN as described for Poisson sampling above. Exponential sampling does not reach the optimal expected overlap in (v) and (vi), but is not too far away, see the numerical examples below. A formula for expected overlap for positive coordination was given in Ohlsson (1996).

Since this procedure is very simple to implement, (i) is fulfilled along with (ii)-(iii); (iv) is not relevant since  $n=1$ . We just mentioned that (v)-(vi) are fulfilled. Finally, it is not hard to see that (vii) is satisfied. We conclude that Exponential sampling is a strong candidate for coordinated PPS sampling when  $n=1$ .

##### 4.2 The case $n>1$

Unlike the  $n=1$  case, PPS sampling with  $n>1$  can be done in several different ways. Most of these can not be used in connection with PRN, since this requires procedures that use individual random numbers for the units. A natural idea is to extend Exponential sampling to  $n>1$ , by selecting the units with the  $n$  smallest transformed random numbers. Unfortunately, this yields so called successive sampling (Cochran, 1977, Section 9A.8) which is not strictly PPS. The actual inclusion probabilities may be quite far from the target values. Cochran (1977) presents several techniques for handling this problem, of which we consider Brewer's method for  $n=2$ . In our context, this method can be applied by drawing the first unit as in Exponential sampling, but with transformed random numbers

$$\xi_i = -\frac{\log(1 - X_i)}{p_i(1 - p_i)/(1 - 2p_i)} \quad (3)$$

After removing the unit drawn in the first round, a second unit is drawn with the transformed random numbers of Exponential sampling  $\xi_i = -\log(1 - X_i) / p_i$ . Cochran (1977, Section 9A.8) shows that these two steps yield a size  $n=2$  sample with the required inclusion probabilities, i.e. meets (ii) and (iii). The procedure is relatively simple (i), and has the same properties as Exponential sampling as regards (v)-(vii). It also allows unbiased variance estimation with the Sen-Yates-Grundy estimator. This is not a single-sum estimator, though, so (iv) is not completely met.

The extension of Brewer's method to  $n>2$  by Sampford (1967) is rather complicated and will not be considered here.

Ohlsson (1990 and 1998) gave an alteration of Poisson sampling called Sequential Poisson sampling (SPS). This procedure uses the transformed random numbers  $\xi_i = X_i / p_i$  and selects the  $n$  units with the smallest such numbers.

The sample will be close to a Poisson sample, since the latter selects the units with  $\xi_i \leq n$ . Not surprising, the properties are approximately the same as those of Poisson sampling, but the size is fixed. The procedure is very simple (i), admits simple-sum variance estimation (iv) and yields independent strata (vii). It is approximately PPS, in the meaning of (ii) and (iii), a fact which is motivated by asymptotics and simulation in Ohlsson (1998). Even though sample coordination is not optimal, in terms of maximum and minimum expected overlap in (v) and (vi), respectively, we can expect the overlap not to be too far from the optimum of Poisson sampling. The case of positive coordination is investigated in a simulation study in the next chapter.

Rosén (1997) gave an alteration of SPS, called *Pareto sampling* (PAS), with transformed random numbers of “odds ratio type”

$$\xi_i = \frac{X_i / (1 - X_i)}{np_i / (1 - np_i)} \quad (4)$$

The properties are similar to those of SPS, but PAS is somewhat closer to the target inclusion probabilities in (ii) and (iii). The closeness to the optimum in (v) is investigated in the simulation study below.

## 5. COMPARISON WITH OTHER PROCEDURES

### 5.1 The case $n=1$

In the literature, there are several (non-PRN) procedures for positive coordination (maximizing overlap) of two PPS size  $n=1$  samples. The pioneering procedure by Keyfitz (1952) assumes the same stratification for both samples. Conditional on the first sample, Keyfitz’ method focuses on the second sample. Suppose unit  $i$  was selected in the first sample. Keyfitz’ method retains this unit in sample if it is increasing, i. e.  $p_i \leq p'_i$ . Else, the unit is retained with probability  $p'_i / p_i$ . If the unit is rejected, one of the increasing units are selected with probability proportional to the increments  $(p'_i - p_i)$ . It is easily verified that this simple procedure is strictly PPS and is optimal in terms of maximizing expected sample overlap, when we have the same stratification for both samples.

Keyfitz’ procedure can be extended to the case with (somewhat) changing strata. We will describe this procedure for a single new stratum. First identify an old stratum which will be considered as the predecessor of our new stratum. Units coming from other old strata, *immigrants*, are treated as births, i.e., they are assigned the value  $p_i = 0$ . Any first sample selection among immigrants is ignored. Then the original Keyfitz algorithm is applied to the new stratum, but the first two steps are only applied to an eventual initial selection that is not an immigrant.

Kish and Scott (1971) note that this procedure can be far from optimal in terms of expected overlap unless we have very small differences in the stratification of the two samples. They provide three methods (beside the extended Keyfitz procedure) for the case with arbitrary stratification of the two samples. We shall consider only Method II, which is claimed by Kish and Scott to give the largest overlap of the three, without being very complicated. The procedure is an elaborated extension of Keyfitz’ method. For a description, we refer to the original article.

The procedure has the disadvantage of distorting the independence between the strata, i.e., it does not fulfil requirement (vii). As already noted, this implies that the procedure can not be applied repeatedly to the same survey. Like Keyfitz’ procedure, Kish and Scott only concerns the second sample, so that (ii) is trivial. In their section 6.2, Kish and Scott (1971) prove that the procedure fulfils (iii). The proof relies on the independence of the initial strata, though. A consequence of the dependence of the new strata is therefore that the procedure is not strictly valid for repeated use, i.e., (iii) is only valid the first time the method is applied. The expected overlap is quite close to optimum, see Section 6.1.

## 5.2 The case $n>1$

Causey, Cox and Ernst (1985) suggested a procedure which maximizes (or minimizes) the expected overlap, subject to the constraints of having the required target probabilities in both samples, i.e. conditions (ii) and (iii). The problem is solved by linear programming methods.

By design, this procedure fulfils our requirement (ii)-(iii) and is optimal in terms of our choice of (v) or (vi). Ernst and Ikeda (1995) note two difficulties with the procedure which can make it unusable in practice. One is that (vii) is not fulfilled, with the same difficulty as for Kish and Scott to apply the procedure repeatedly for the same survey, especially if  $n>1$ . The second difficulty is that the transportation problem may be too large to solve in practice. In any case, (i) is not fulfilled.

Ernst (1999) reviews several alterations of the Causey, Cox and Ernst procedure, non of which fulfills all our requirements.

Sunter (1989) presents an interesting procedure that is applicable for maximizing overlap of two samples with any sample size  $n$ . It is a generalization of Keyfitz' procedure. Like the latter, Sunter's procedure was not primarily designed for handling stratum changes and can be expected to give an overlap far from maximum when we have large stratum differences between the two samples.

## 6. NUMERICAL EXAMPLES

Below we report results from two numerical studies, one for  $n=1$  and one for  $n=2$  to 4.

### 6.1 The case $n=1$

The first study concerns expected overlap of different procedures for maximizing overlap in the case  $n=1$ . It is based on the so called MU284 population of 284 Swedish municipalities presented in Appendix B of Särndal et al. (1992). See <http://statlib@lib.stat.cmu.edu/datasets/mu284>.

Draw probabilities are either equal or proportional to the number of inhabitants, for the first sample we use figures from 1975 (P75), and for the second sample those of 1985 (P85). First sample strata are either defined by the regional REG variable, giving 8 strata of sizes between 15 and 56, or by the CL variable, with 50 small strata with sizes between 5 and 8.

The expected overlap was computed exactly (no simulation) for PRN Exponential sampling (EXP), the extended Keyfitz (KEY), the Kish and Scott (K&S), and the Cox-Causey-Ernst (CCE) procedure.

As benchmarks we use (a) the case without any overlap control, with both samples drawn independently (IND) and (b) the non-achievable upper limit (2), added over the strata. For further details on the study, see Ohlsson (1996).

*Table 0. MU284 population. Expected overlap in eight regional (REG) strata. Unequal probabilities. 50% of the units move from each stratum to an adjacent stratum.*

Stratum	$N_h$	IND	EXP	KEY	K&S	CCE	Limit
1	36	0.040	0.524	0.431	0.592	0.592	0.718
2	37	0.183	0.772	0.588	0.818	0.874	0.997
3	35	0.084	0.689	0.554	0.793	0.793	0.942
4	35	0.042	0.614	0.488	0.703	0.703	0.845
5	48	0.030	0.578	0.446	0.676	0.676	0.764
6	49	0.105	0.700	0.581	0.814	0.814	0.916
7	21	0.089	0.677	0.609	0.763	0.769	0.948
8	23	0.109	0.645	0.552	0.716	0.741	0.928
Sum	284	0.682	5.200	4.249	5.875	5.964	7.058
Percent		9	65	53	73	75	88

For the remaining set-ups, we present no stratum details but just the sum in percent of the number of sampled units.

Table 1.  $n=1$ . MU284 population. Expected overlap in percent of total sample size. No changes in strata.

Strata	Prob	IND	EXP	KEY	K&S	CCE	Limit
Medium (REG)	Unequal	9	96	97	97	97	97
	Equal	1	100	100	100	100	100
Small (CL)	Unequal	30	97	98	98	98	98

Table 2.  $n=1$ . MU284 population. Expected overlap in percent of total sample size. One unit changes stratum.

Strata	Prob	IND	EXP	KEY	K&S	CCE	Limit
Medium (REG)	Unequal	9	92	94	94	95	96
	Equal	3	95	97	97	97	99
Small (CL)	Unequal	29	78	80	80	84	87

Table 3.  $n=1$ . MU284 population. Expected overlap in percent of total sample size. 50% of units change stratum.

Strata	Prob	IND	EXP	KEY	K&S	CCE	Limit
Medium (REG)	Unequal	9	65	53	73	75	88
	Equal	3	66	52	75	75	89
Small (CL)	Unequal	26	64	48	69	72	81

Table 4.  $n=1$ . MU284 population. Expected overlap in percent of total sample size. One third of the units move to the next stratum, one sixth to the following.

Strata	Prob	IND	EXP	KEY	K&S	CCE	Limit
Medium (REG)	Unequal	7	59	46	70	-	83
	Equal	3	61	42	69	-	88
Small (CL)	Unequal	-	-	-	-	-	-

Note: CCE (Causey-Cox-Ernst procedure) intractable in two strata. Unequal case not treated.

## 6.2 The case $n>1$

For  $n=2, 3, 4$  we have conducted a simulation study of expected overlap for sequential Poisson sampling (SEP), and Pareto sampling (PAR). For  $n=2$ , we also consider exponential sampling (EXP), in the form using Brewer's recalculated draw probabilities. The benchmarks IND and Limit are as in the preceding section.

We use data from a master frame of the Swedish National Road Administration. The statistical units are stretches of road and the size measure is derived from traffic mileage per year. The units are stratified according to region and type of road, altogether 28 strata with 2523 units. We let 10% of the units change strata between the two sampling occasions. A more detailed report of the study is given in Ohlsson (1999). The data are available at the address <http://www.matematik.su.se/~esbj/roads.dat>. The simulations were run with 14000 iterations.

Table 5.  $n=2$ . Road population. Expected overlap by stratum.

Stratum no.	$N_h$	IND	EXP	SEP	PAR	Limit
11	44	0.12	1.77	1.77	1.77	1.97
12	41	0.13	1.66	1.66	1.66	1.78
13	58	0.12	1.80	1.79	1.79	1.93
14	292	0.03	1.75	1.74	1.74	1.93
21	35	0.14	1.78	1.78	1.78	1.94
22	80	0.07	1.73	1.72	1.72	1.91
23	82	0.08	1.72	1.72	1.72	1.91
24	304	0.03	1.73	1.73	1.73	1.91
31	12	0.49	1.77	1.77	1.78	1.93
32	6	0.63	1.61	1.60	1.61	1.77
33	34	0.19	1.56	1.54	1.55	1.79
34	61	0.11	1.77	1.76	1.77	1.96
41	39	0.16	1.72	1.72	1.73	1.95
42	79	0.07	1.70	1.70	1.70	1.88
43	73	0.08	1.69	1.69	1.69	1.91
44	318	0.02	1.73	1.73	1.73	1.94
51	30	0.16	1.70	1.69	1.70	1.95
52	56	0.08	1.70	1.70	1.70	1.89
53	44	0.12	1.66	1.66	1.66	1.91
54	170	0.04	1.67	1.66	1.66	1.92
61	37	0.16	1.73	1.73	1.73	1.90
62	68	0.08	1.75	1.75	1.75	1.96
63	53	0.10	1.66	1.65	1.66	1.92
64	301	0.02	1.65	1.64	1.64	1.86
71	21	0.25	1.66	1.66	1.66	1.86
72	34	0.15	1.65	1.64	1.65	1.94
73	29	0.17	1.63	1.62	1.63	1.93
74	122	0.05	1.71	1.71	1.71	1.92
<i>Sum</i>	<i>2523</i>	<i>3.9</i>	<i>47.6</i>	<i>47.5</i>	<i>47.6</i>	<i>53.4</i>

For  $n=3$  and 4 we only give the sum over strata.

Table 6. Road population. Expected overlap, aggregated over strata. 10% of units change stratum.

Sample size	IND	EXP	SEP	PAR	Limit	$n$
$n=2$	3.9	47.6	47.5	47.6	53.4	56
$n=3$	8.7	-	72.8	72.9	80.1	84
$n=4$	15.5	-	98.6	98.7	106.7	112

### 6.3 Inclusion probabilities

The PRN techniques for  $n>1$  are only approximately (asymptotically) PPS. Numerical studies indicate that the approximation is very good in many situations. Ohlsson (1990, 1998) reports a simulation study on CPI data, where SEP is very close to being unbiased. Rosén (1998) studies exact actual inclusion probabilities (AIP) for SEP and PAR in an artificial, but nevertheless interesting, situation, viz. when all units have the same inclusion probability except for one odd unit. Even with sample sizes as small as  $n=2$ , the relative error in the AIP for PAR is never larger than 2% and in most cases it is much smaller. SEP is a bit further from the target probability. The AIPs were also computed in the simulation study mentioned in Section 6.2. The results are reported in Ohlsson (1999). Here again PAR is quite close to the unbiased case, with SEP performing a little bit less good.

## 7. CONCLUSIONS

We first consider the problem of *maximizing* overlap, which is the concern of the numerical studies. Using any of the procedures under consideration gives a great increase in (expected) sample overlap, as opposed to drawing independent samples (IND). For  $n=1$ , the Kish and Scott procedure is quite close to the optimal (in this respect) Cox, Causey and Ernst procedure. When there are great differences between the strata of the two samples, Keyfitz' procedure is rather far from the maximum expected overlap. Exponential sampling is a bit away from the optimum, but much less so than Keyfitz. Since the K&S and CCE both suffer from the problem of dependent strata, i.e. violate (vii), and since the former does not fulfill (vii) and the latter violates (i), we consider Exponential sampling a good compromise in the search for a procedure that fulfills (i)-(iii) and (v)-(vii) in as far as possible. Of the mentioned procedures, Exponential sampling and CCE are the only ones that can be used for *minimizing* overlap (vi).

Turning to the case  $n>1$ , we concluded in Section 5.2 that CCE can be unusable in practice and that Sunter's procedure can be expected to give a low sample overlap when there are great differences in stratification between the two samples. This leaves us with the PRN procedures, which are all equal in sample overlap in our simulation studies. Since PAR is a bit closer to the right inclusion probabilities it is generally preferable to SEP. EXP is an alternative for  $n=2$ , being strictly unbiased, but it suffers from a more complicated variance estimation procedure.

In any case, the various PRN techniques investigated here are simple and efficient for simultaneous negative and positive sample coordination of any number of surveys, with any draw probabilities and any stratification. When  $n>1$ , they all allow for variance estimation, and all but Exponential have an associated single-sum variance estimator. In summary, the PRN techniques fulfill (i)-(vii) even though they do not strictly maximize/minimize the expected sample overlap.

The overall conclusion is that PRN procedures, n.b. Exponential sampling for  $n=1$  and maybe  $n=2$ , and Pareto sampling for  $n>1$ , are competitive as procedures for controlling sample overlap.

## 8. REFERENCES

- Aires, N. (1999). "Comparisons Between Conditional Poisson Sampling and Pareto  $\pi$ ps Sampling Designs," Contributed paper, Bulletin of the International Statistical Institute, 52nd Session.
- Ålenius, M. (1990). "Storleksstratifiering eller pps-urval? En jämförande studie för KPI-data," *F-METOD NR 25, Statistics Sweden*. (In Swedish.)
- Bailey, J. T. and Kott, P.S. (1997). "An Application of Multiple List Frame Sampling for Multi-Purpose Surveys," *ASA Proceedings of the Section on Survey Research Methods*.
- Brewer, K.R.W., Early, L.J. and Joyce, S.F. (1972). "Selecting several samples from a single population," *Australian Journal of Statistics*, 14, 231-239.
- Brewer, K.R.W., and Hanif, M. (1983). *Sampling With Unequal Probabilities*. Springer, New York.
- Causey, B. D., Cox, L. H. and Ernst, L. R. (1985), "Application of Transportation Theory to Statistical Problems," *Journal of the American Statistical Association*, 80, 903-909.
- Cochran, W. G. (1977). *Sampling Techniques 3d ed.*, New York: John Wiley.
- Ernst, L. R. (1999), "The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results," Invited Paper, Bulletin of the International Statistical Institute, 52nd Session.
- Ernst, L. R. and Ikeda, M. M. (1995), "A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys," *Survey Methodology*, 21, 147-157.

- Keyfitz, N. (1951), "Sampling with Probabilities Proportional to Size," *Journal of the American Statistical Association*, 46, 105-109.
- Kish, L. and Scott, A. (1971), "Retaining Units After Changing Strata and Probabilities," *Journal of the American Statistical Association*, 66, 461-470.
- McKenzie, R. and Gross, B. (1999), Synchronized Sampling at the Australian Bureau of Statistics. In this volume.
- Ohlsson, E. (1990), *Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer Price Index*, R & D Report 1990:6, Statistics Sweden.
- Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers," In *Business Survey Methods*, New York: Wiley, 153-169.
- Ohlsson, E. (1996), *Methods for PPS Size One Sample Coordination*, Research Report No. 194, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University.
- Ohlsson, E. (1998), "Sequential Poisson Sampling," *Journal of Official Statistics*, 14, 149-162.
- Ohlsson, E. (1999), *Methods for PPS Size One Sample Coordination*, Research Report No. 210, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University.
- Rosén, B. (1997b), "On Sampling with Probability Proportional to Size," *Journal of Statistical Planning and Inference*, 62, 159-191.
- Rosén, B. (1998), On Inclusion Probabilities for Order Sampling, R & D Report 1998:2, Statistics Sweden. To appear in *Journal of Statistical Planning and Inference*?
- Royce, D. (1999), "Issues in Co-ordinated Sampling at Statistics Canada," In this volume.
- Sampford, M. R., (1967), "On Sampling Without Replacement with Unequal Probabilities of Selection," *Biometrika*, 54, 499-513.
- Särndal, C. E., (1996), "Efficient Estimators With Simple Variance in Unequal Probability Sampling," *Journal of the American Statistical Association*, 91, 1289-1300.
- Särndal, C. E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- Srinath, K.P. and Carpenter, R.M. (1995). "Sampling Methods for Repeated Business Surveys," In *Business Survey Methods*, New York: Wiley, 153-169.
- Sunter, A. B (1989), Updating Size Measures in a PPSWOR Design, *Survey Methodology*, 15, 253-260.

## DISCUSSION OF SESSION 31: COORDINATING SAMPLING BETWEEN AND WITHIN SURVEYS

Lawrence R. Ernst, U.S. Bureau of Labor Statistics  
BLS, 2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212-0001, U.S.A.  
ernst\_l@bls.gov

### 1. CATEGORIZATION OF THE THREE PAPERS

Although there are three papers in this session, I view them as fitting into two categories of procedures for controlling sample overlap. The major focus of Ohlsson's paper is on an overlap procedure that he developed, exponential sampling, which is in a class of procedures most commonly used in selecting PSUs for household surveys, although some establishment surveys also use PSUs. Perhaps the key characteristic of such procedures, which originated with Keyfitz (1951), is that the sample size per stratum,  $n$ , is always small, typically either 1 or 2, and we consequently designate these procedures as S procedures. The following are common characteristics of S procedures. For a given stratum,  $n$  is predetermined, although in some designs  $n$  may vary by stratum. Sample units are selected pps. If  $n > 1$ , the joint selection probabilities within a stratum are predetermined. Most S procedures, including Ohlsson's procedure, although not all, allow for different stratifications in the designs being overlapped. These procedures have been developed for overlap maximization and/or minimization, but not partial rotation. S procedures generally do not use PRNs, with Ohlsson's procedure the only exception that I am aware of among procedures that strictly preserve the desired selection probabilities. (We restrict the discussion to such procedures.) Finally, S procedures typically overlap only two samples at a time, with again Ohlsson's procedure an exception.

The McKenzie and Gross (M&G) and Royce papers, in contrast, discuss overlap procedures typically used for selecting establishments from a stratified list frame, with a key characteristic that these procedures must be capable of overlapping samples for which  $n$  is large. Consequently, we designate these procedures as L procedures. In L procedures,  $n$  for a stratum may be either predetermined, as in synchronized sampling of the M&G paper, or variable, as in all the Statistics Canada (STC) procedures described in the Royce paper. The selection of the sample units is commonly, as in both of these papers, although not exclusively, with equal probability. Joint selection probabilities are generally neither predetermined, nor calculated. Typically L procedures do not attempt to control overlap for units changing strata. This is because, in addition to the extra complexity of attempting to do so, the most common stratification change is the occasional establishment changing size class. Both of these papers are exceptions because they do discuss procedures that control overlap with restratification. In fact, this is a key issue in the M&G paper. L procedures are used for overlap maximization, minimization, and partial overlap, and all three applications are discussed in both of these papers. L procedures commonly use PRNs, with the procedure used by the Canadian Monthly Wholesale and Retail Trade Survey the only exception in these two papers. Some L procedures are applicable to the overlap of more than two surveys at a time.

### 2. OHLSSON PAPER

I consider the procedure described in Ohlsson's paper to be highly innovative and an extremely important contribution to controlling overlap. Before discussing other details of the features of this procedure, let me mention one important feature that is not mentioned in this paper, but is described in Ohlsson (1996). Suppose an initial sample has been chosen, that is one not overlapped with a previous sample, without using Ohlsson's procedure. Although it might appear to be too late then to overlap a subsequent sample with this initial sample using his procedure, it is shown in Ohlsson (1996) that for  $n = 1$ , PRNs can be assigned retrospectively, conditioned on the initial sample, and the Ohlsson procedure then applied to subsequent samples with all the properties of the procedure remaining unchanged. It is not presently known whether this result can be extended to  $n > 1$ .

Ohlsson's procedure, like all overlap procedures, has both advantages and disadvantages. The short list of disadvantages will be mentioned first. It is the only S overlap procedure that I am aware of that in the case when  $n = 1$  and two samples with identical stratification are overlapped does not yield the optimal overlap. All other S procedures reduce to Keyfitz's (1951) procedure in that case. In particular, Keyfitz's procedure always retains in the new sample any sample unit in the initial design that has a selection probability at least as large in the new design as in the initial design. However, with Ohlsson's procedure, such a unit can be replaced in the new sample by a unit with a selection probability that has increased by a larger percentage in the new design. In addition, there are

limitations on the use of this procedure. If  $n > 1$  and Ohlsson's procedure was not used in selecting the initial sample, then it cannot be used to select subsequent samples since the procedure for retrospectively assigning PRNs mentioned above has only been developed for the case  $n = 1$ . In addition, regardless of  $n$ , if another overlap procedure had previously been used that destroyed the independence of sampling from stratum to stratum, then Ohlsson's procedure cannot be used.

Most of the advantages of Ohlsson's procedure particularly apply when the surveys being overlapped have different stratifications, so we will confine our discussion to this case, which is the more common case in practice when  $n$  is small. Ohlsson's procedure is quite simple to implement. This is particularly noteworthy when  $n > 1$ , since most alternative procedures, including Causey, et al. (1985), Ernst (1986), and Ernst and Ikeda (1995), employ linear programming algorithms, which are generally not simple to implement. Also, which is the key point of the procedure, it preserves the independence of sampling from stratum to stratum. Overlap procedures that require the same stratifications in the surveys overlapped also automatically satisfy this independence. Besides these procedures, the only other procedures that I am aware of that preserve this independence do not predetermine the sample size and hence are not S procedures. In addition to the advantages in variance estimation, Ohlsson's procedure can be applied repeatedly as a result of this independence property. Among alternative procedures, some, such as Kish and Scott (1971), Causey, et al. (1985), and Ernst and Ikeda (1995), can only be used once, since they assume this independence; while others, such as Perkins (1970) and Ernst (1986), which can be used repeatedly, tend to yield an overlap that is further from optimal. To illustrate the resulting problems, consider the selection of the PSUs for the U.S. Census Bureau's Survey of Income and Program Participation. This survey is redesigned every 10 years. The PSUs were selected independently in the 1980s redesign. For the 1990s selection, the sample was overlapped with the 1980s sample using the Ernst and Ikeda (1995) procedure. For the 2000s redesign, since the Ernst and Ikeda procedure cannot be used again, the Ernst (1986) procedure will be used, which should not produce as large an overlap. It might have been better to have used Ohlsson's procedure throughout. It would be interesting to empirically compare the overlap produced by Ohlsson's procedure to that produced by the Ernst (1986) procedure, which produces the best overlap among those procedures that can be used repeatedly because they do not assume independence from stratum to stratum. I suspect that the Ernst procedure would be superior when the surveys overlapped have similar stratifications, since the Ohlsson procedure is not optimal when the stratifications are identical. However, when the stratifications are very different, Ohlsson's procedure may be superior, since it is known that the Ernst procedure does not generally produce an overlap that is close to optimal under this condition.

### 3. MCKENZIE AND GROSS PAPER

I have always been impressed with the general approach in synchronized sampling of moving the endpoints to the right while keeping the sampling size fixed, which prevents units that leave the sample at one time period as a result of more births than deaths from reentering the next time period as a result of more deaths than births.

The focus in this paper, however, is not on synchronized sampling in general, but on maximizing or minimizing overlap with another stratification of the same survey or with one or more other surveys. As the authors note, the overlap attained using their procedure is far from optimal because of the "partial intervals" problem. The only way that I am aware of to avoid this problem without a transformation of the PRNs is to have the new sample consist of several partial intervals with only units in each partial interval from the corresponding old stratum included, if possible, rather than a single interval with all units included. I suspect that this solution would cause too many operational problems to be seriously considered.

The fact that there is a "bias" with their procedure when the starting point of the selection interval is the last point that produces the maximum score is illustrated by the following simple example. Suppose a stratum in the new design consists of two complete old strata, A and B, with  $n = 1$  for the new stratum and both old strata, and with A consisting of more units than B. Then the probability is .5 that the last point that produces the maximum score is the sample point in stratum A and the probability is .5 that it is the sample point in stratum B. Consequently, the selection of the sample unit in the new stratum would not be with equal probability, but instead each unit that was in stratum B would have a higher selection probability than each unit that was in stratum A. I agree, as mentioned in M&G, that the use of the point after last occurrence of maximum score rather than the point of the last occurrence as the starting point of the selection interval tends to reduce the size of the misallocation, although I believe this whole issue is complex and needs further study. Furthermore, for very small  $n$ , particularly for  $n = 1$ , this choice of

starting point may reduce the effectiveness of the overlap. To illustrate, consider the above example with the point after the last occurrence of the maximum score as the starting point. Then the sample unit in the new stratum would only have been a sample unit in the old design if the stratum A sample unit and stratum B sample unit are the first and last points the new stratum, an event with probability lower than the probability of overlap when the sample unit in the new stratum is selected independently of the selection of the sample units in the old strata.

The avoidable load measure used in the empirical study is interesting. However, it is a measure of high load, while the maximum score used in the overlap procedure attempts to minimize a somewhat different measure. For example, selecting two units, one with a high load and one with a low load may result in a lower score but a higher contribution to the avoidable load than selecting two units with medium loads. I believe it also would be worthwhile to compare the avoidable load obtained in the empirical study using their overlap procedure to the avoidable load obtained selecting the samples for these 12 surveys independently.

#### **4. ROYCE PAPER**

Although several approaches for sample coordination are discussed in this paper, STC appears to be standardizing for within survey sample coordination around GSAM, which uses collocated random numbers (CRNs) that have at least two advantages over standard PRNs due to the equal spacing of the selection numbers. First, CRNs help reduce the variability of the sample size when using a fixed selection interval, although, particularly because of births and deaths, they do not completely eliminate it. In addition, an attempt to minimize overlap among surveys by assigning each survey different starting points an appropriate distance apart cannot fail with CRNs, as it can with PRNs if most of the PRNs are clustered close to each other. CRNs also have disadvantages in comparison with PRNs. CRNs cannot be assigned on a flow basis, which is why they could not be used in the Tax Estimates Program. Also coordinating surveys with different stratifications or when restratifying can be more complicated with CRNs, as noted by Ohlsson (1995), since the CRN for a unit is a function of the number of units in the stratum.

One of the important features of GSAM is a procedure for maximizing overlap when restratifying a survey. GSAM accomplishes this while avoiding the partial intervals problem discussed in the M&G paper and thus generally produces a larger overlap. This is done by assigning new selection numbers to the units in a new stratum in a manner that clusters together as much as possible the units that were in sample under the old stratification. Using this approach for coordinating among different surveys might lead to complications, however, because the same unit would have different selection numbers for different surveys.

An alternative to the procedure described for minimizing the overlap between the crops and livestock surveys would be to choose appropriate starting points for the selection intervals for the two surveys and move both intervals to the right. For example, if the sample for one survey was three times larger than the sample for the other in a stratum, then the selection interval for the larger sample could start at 0 and for the smaller sample at 0.75. This alternative should result in a longer time period before the samples could overlap. However, I understand that this alternative was among those considered by STC, but that it produced a larger overlap than the procedure adopted.

The procedure described for reducing respondent burden in SEPH for multi-establishment enterprises results in biased estimates. In particular, I believe there may be a large underallocation of establishments in large enterprises because of this procedure.

The network sampling approach used in UES is quite interesting. It complicates estimation and variance estimation, however, as discussed in Simard and Hidioglou (1999).

#### **5. REFERENCES NOT LISTED IN SESSION PAPERS**

- Ernst, L. R. (1986), "Maximizing the Overlap Between Surveys When Information Is Incomplete," *European Journal of Operational Research*, **27**, pp. 192-200.
- Perkins, W. M. (1970), "1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata," memorandum to Joseph Waksberg, Washington, DC: U.S. Bureau of the Census.

*Any opinions expressed in this discussion are those of the author and do not constitute policy of BLS.*

