PPES SAMPLING OF TWO SUBDOMAINS WITH INDEPENDENT PROBABILITIES

Jesse Pollock, U.S. Bureau of the Census

I. Introduction:

When sampling from a frame where values for some correlated auxiliary variable are known for for each unit of observation, the sampler has a choice of how to use the auxiliary variable. Reduction of sampling error is of primary importance but ease of sampling, processing, and administering the survey may also also influence the decision. Two alternatives which are often used are:

1. Take a simple random sample and use a ratio estimator.
2. Sample with probability proportional to the auxiliary variable and use the Horvitz-Thompson estimator.

In general practice, it is difficult to choose between these two with regard to smaller variance, unless the form of the regression of the response on the auxiliary variable is known. Suppose one is interested in the situation where there are two domains of interest. Examples are:

1. Yields are to be estimated for each of two crops and acreage planted for each crop is known for each farm in a particular area.
2. Retail receipts for men's and women's shoes are to be individually estimated for shoe stores which sell both and prior year receipts are available for each store and each subdomain (men's and women's).

Sampling proportional to a single measure may present difficulities. If a weighted average of the auxiliary variables is used as a measure of size, the sampling variance may increase substantially, particularly for one of the variables. If two independent samples are chosen, the costs may increase substantially, since the number of units common to both samples will be relatively small.

In this paper, the feasibility of using conditional probabilities to select the second sample which maximize the expected number of units common to both samples are examined. All sample selection is assumed to be without replacement. The methods and programs for sample selection are described in Section II. In Section III, comparisons are made to ratio estimation for simple random samples drawn without replacement from four synthetic universes, all of size 365. In Section IV, observations from the study are noted and paths for further research are discussed.

II. Description of the Procedure:

In Causey, Cox, and Ernst (1983), a general solution is presented for maximizing the overlap between two samples. The solution uses general linear programming algorithms. For a design with a large number of strata, where very few units (perhaps 1 or 2) are being chosen per stratum, and particularly when stratum definitions change, this procedure is quite efficient. However, it is necessary to formulate every possible first sample and its probability of selection and every possible (or desired) second sample and its probability of selection consistent with predetermined selection probabilities (i.e., sizes) for each unit. For large universes and sample sizes these methods may become impractical.

In the case being discussed here, there are two important restrictions which can simplify computation.

1. The units (clusters) in the frame are identical (e.g., corn and soybean yields are to be estimated for the same farms).
2. Systematic PPES sampling is used for the first sample.

The importance of the second condition is that for any sample size, n, there are at most N (the universe size) possible systematic samples (see Hartley (1966)).

It is still necessary to formulate every possible first sample and its selection probability, but since it is restricted to size N, moderate size universes (or strata) can be accommodated.

As stated above, the first step is to identify all possible systematic PPES samples for the first domain. These samples are indexed by j, j=1, ...N. The choice of "first variable" is arbitrary unless there is a natural order. For example, if selection is for the same characteristic in two time periods, "first" is obviously determined by the time period. For each unit, i, an initial conditional probability, for selection in the second sample, conditional on first sample j, P_{ij}, is assigned as follows:

1. If $p_{2i} \geq p_{1i}$

$P_{ij} = 1$ when the i^{th} unit is in first Sample j

$$= \frac{p_{2i} - p_{1i}}{1 - p_{1i}} \quad \text{if not selected}$$

2. If $p_{2i} < p_{1i}$

$$P_{ij} = \frac{p_{2i}}{p_{1i}} \text{ when the } i^{th} \text{ unit is in first Sample j}$$

$$= 0 \quad \text{if not selected}$$

where p_{1i}, p_{2i} are the predetermined probabilities of selecting unit i in the first and second samples, respectively. In either case, the expected second selection probability is p_{2i} and the expected number of units common to both samples will be maximal.[1/] The problem here is that the expected second sample size $\sum_{i}^{N} P_{ij}$ is not fixed for any given first sample, though the expected size over all possible first samples is $\sum_{i=1}^{N} p_{i2} = n_2$. [The examples presented in the next section all use $n_1 = n_2$, though this is not necessary. The flexibility of sample size may, in fact, be an advantage of this procedure over the ratio estimation procedure or any single sample procedure but is not explored in

this study.] It should be noted that if the second sample is large enough so that a small variation in the sample size is tolerable, these initial conditional probabilities are sufficient for selecting the second sample. They are very simple to calculate, since they depend only on the first sample actually drawn, rather than all possible first samples, and so can be easily used for any type of PPES sampling.

To comply with the further restriction of a constant second sample size, the matrix of initial conditional probabilities P_{ij} (a column is a given first sample, a row is a unit in the frame) is converted to a matrix of joint probabilities of selecting first sample j, and unit i in the second sample. $P'_{ij} = P_{ij}\pi_j$ where π_j is the probability of selecting first sample j. Proportional iterative allocation (raking procedure) is used to find find a matrix of joint probabilities P''_{ij} that satisfy the following:

(1) $\sum_i P''_{ij} = p_{2i}$ (Horvitz Thompson estimation for variable 2 using p_{2i} is design unbiased)

(2) $\sum_i P''_{ij} = n_2 \pi_j$ (constant sample size n_2)

(3) $P''_{ij} = P_{ij} = \pi_j$ when $P_{ij} = 1$ (conditional certainty units remain certainty)

(4) $P''_{ij} = 0$ when $P_{ij} = 0$

Condition (3) is easily accomplished by substituting 0 for P''_{ij} whenever $P''_{ij} = \pi_j$ and subtracting π_j from the appropriate marginals. The raking procedure multiplies each entry in column j by $n'_2 \pi_j / \sum_i P^{(m_r)}_{ij}$ (where $n'_2 = n_2 - C_j$). C_j is the number of conditional certainty units in column j and $P^{(m_r)}_{ij}$ is the value of the joint probability after m row iterations; $P^{(m_c)}_{ij}$, the joint probability after m column iteration) and then each row by

$$p'_{2i} / \sum_j P^{(m_c)}_{ij} \quad \text{where } p'_{2i} = p_{2i} \text{ if } p_{2i} < p_{1i}$$

$$= p_{2i} - p_{1i}, \text{ otherwise.}$$

This is done iteratively until the matrix converges to some predetermined level. Conditions (3) and (4) will ensure maximum overlap, since the raking will then redistribute the joint probability within overlap cells (for $p_{1i} > p_{2i}$) or nonoverlap cells (for $p_{1i} < p_{2i}$), only. In practice, three or four iterations are almost always sufficient. Conditions for convergence are discussed in Greenberg and Fagan (1984).[2] The second sample can then be selected with the resultant probabilities.

The following simplified example illustrates the procedure. Here, $N = 6$, $n_1 = n_2 = 3$; the units will be referred to as A_i with the following selection probabilities:

Unit	p_1	p_2
A_1	.8	.6
A_2	.7	.4
A_3	.4	.5
A_4	.5	.6
A_5	.4	.5
A_6	.2	.4

Taking systematic samples with p_{1i}, there are five possible first samples.

	Sample			Probability (π_j)
S_1	A_1	A_2	A_4	.4
S_2	A_1	A_2	A_5	.1
S_3	A_1	A_3	A_5	.3
S_4	A_2	A_3	A_6	.1
S_5	A_2	A_4	A_6	.1

The initial conditional probabilities are then:

unit(samples)	selected	not selected
A_1 ($S_1S_2S_3$)	.750	0
A_2 ($S_1S_2S_4S_5$)	.571	0
A_3 (S_3S_4)	1.000	.167
A_4 (S_1S_5)	1.000	.200
A_5 (S_2S_3)	1.000	.167
A_6 (S_4S_5)	1.000	.250

This can be expressed in the matrix of joint probabilities as:

unit	p_{2i}	$n_2\pi_j$:	S_1 1.2	S_2 .3	S_3 .9	S_4 .3	S_5 .3
A_1	.6		.300	.075	.225	0	0
A_2	.4		.229	.057	0	.057	.057
A_3	.5		.067	.017	.300	.100	.017
A_4	.6		.400	.020	.060	.020	.100
A_5	.5		.067	.100	.300	.017	.017
A_6	.4		.100	.025	.075	.100	.100

Note the respective expected sample sizes are 2.91, 2.94, 3.20, 2.94 and 2.91.

After replacing certainty cases by zero, we have:

		S_1 .8	S_2 .2	S_3 .3	S_4 .1	S_5 .1
A_1	.6	.300	.075	.225	0	0
A_2	.4	.229	.057	0	.057	.057
A_3	.1	.067	.017	0	0	.017
A_4	.1	0	.020	.060	.020	0
A_5	.1	.067	0	0	.017	.017
A_6	.2	.100	.025	.075	0	0

Four iterations of proportional allocation (two across rows, two across columns) and replacing certainty probabilities yields the final solution.

		S_1 1.2	S_2 .3	S_3 .9	S_4 .3	S_5 .3
A_1	.6	.334	.080	.186	0	0
A_2	.4	.226	.054	0	.060	.060
A_3	.5	.064	.016	.300	.100	.020
A_4	.6	.400	.023	.053	.024	.100
A_5	.5	.064	.100	.300	.016	.020
A_6	.4	.112	.027	.061	.100	.100

or, in terms of selection probabilities:

	S_1	S_2	S_3	S_4	S_5
A_1	.835	.800	.620	0	0
A_2	.565	.540	0	.600	.600
A_3	.160	.160	1.000	1.000	.200
A_4	1.000	.230	.177	.240	1.000
A_5	.160	1.000	1.000	.160	.200
A_6	.280	.270	.203	1.000	1.000

The expected overlap is 2.5 units.

III. Relative Efficiency:

In this section, the CV's for four artificially generated populations of size 365 will be compared for fixed costs, between ratio estimates from a single simple random sample without replacement and the dual PPES (without replacement) samples with maximum overlap. These four populations were constructed so that PPES sampling would be efficient, relative to simple random sampling, for different correlations among the measures and the two response variables. The objective is to see the effect of the dual sample method on the CV's of the subdomains.

First, 365 random numbers between 1 and 2,000 were generated for measures of size of the first subdomain, M_{1i} (i=1,...,365). Then a random error, a_i, independent of M_{1i} was generated for each unit each unit such that $E(a_i) = 0$, and scaled so that ρ_o (M_{1i}, $M_{2i} = M_{1i} + a_i$) $\doteq$.96, where M_{2i} is then the measure of size for the second subdomain. To meet the criteria for efficient PPES samples, values e_{1i} were generated such that $Y_{1i} = M_{1i} + e_{1i} M_{1i}$ [Y_1 is the first response variable]; and $E(e_{1i}/M_{1i}) \doteq 0$. The e_{1i} were scaled so that so that $E(\sum e_{1i}^2) \doteq .024$, i.e. $\rho(M_{1i}, Y_{1i}) = \rho_1 \doteq .99$ and $E(e_i^2)$ was not a function of the M_{1i}. Similarly, e_{2i} was generated so that $\rho_2 \doteq .99$.

These conditions on e_{1i} and e_{2i} yielded a universe where PPES sampling would be preferred for each domain, individually. (See Cochran, Section 9.12). The condition on a_i (i.e. ρ_o = .96) ensured high maximum overlap. This set of 365 x 4 [M_{1i}, M_{2i}, Y_{1i}, Y_{2i}] values is referred to as U1 in Tables 2a and 2b below. U2, U3 and U4 were then generated from U1 by increasing the scaling factors used to generate e_{1i} and e_{2i} (decreasing ρ_1 and ρ_2) and also increasing the scaling factors for a_i to achieve decreasing values of ρ_o, or, in turn, the maximum overlap.

One of the important objectives of this study was to test computing feasibility of the procedure. The programs were written for a mini-computer which dictated the population and sample sizes studied. From each population, all samples of sizes 10 through 50, in increments of 5, were selected systematically, proportional to M_{1i}. Estimates were computed for each sample and first stage CV's for each sample size. After conditional probabilities were computed, as described in the previous section, systematic samples were selected and CV's computed for Y_2. Using the conditional selection probabilities, the second samples were taken systematically using the same order as in the first sample. This order was the same in all four populations. Also, the expected overlap was computed for each sample size. Since, for a particular order, the variances for PPES without replacement sampling are not directly proportional to the inverse of the sample size, the CV's were smoothed over the given range of 10-50. In a few cases, in fact, the CV for a larger sample was greater than for a smaller sample. The values given in Tables 2a and 2b are interpolated from the smoothed curves obtained. Since only a single order was used throughout the study, the smoothing of CV's was used in place of averaging the CV's over all possible orders. The CV's for the simple random sample were computed by standard formulas which, of course, are not dependent on any particular ordering of the units. CV's for ratio estimates were computed without regard to ratio estimation bias, which is assumed to be negligible.

Since the situation being discussed involves cluster sampling, variable survey costs can be put into two components, contacting the cluster (by personal interviewer, mail or telephone) and enumerating the individual elements within that cluster. The second component may involve subsampling, as well. Hence, for a single sample, the cost can be expressed as:

$C_R = n (\bar{C}_1 + 2\bar{C}_2)$, where:
- n : number of clusters sampled
- $\bar{C}_1$: average first component (cluster) cost
- $\bar{C}_2$: average second component (element) cost per cluster for each domain.

We assume equal costs for enumerating each domain. In the dual sample case, we have

$$C_p = n_o (\bar{C}_1 + 2\bar{C}_2) + 2n_1 (\bar{C}_1 + \bar{C}_2)$$

where
- n_o: number of clusters common to both samples
- n_1: number of clusters in sample 1, only.

We assume each sample has the same number (n_1) of clusters unique to that sample.

Hence, for the same cost, $C_R = C_p$, we have $n_o (\bar{C}_1 + 2\bar{C}_2) + 2n_1 (\bar{C}_1 + \bar{C}_2) = n (\bar{C}_1 + 2\bar{C}_2)$

Letting $k = 2\bar{C}_2/\bar{C}_1$ and $r = n_1/n_o$, we have:

$$n_o = n / [1 + r(2+k) / (1+k)]$$

This form for n_o is convenient since r is explicitly calculated for each population. Note that k is the ratio of average second stage costs to first stage costs where both variables are enummerated. One other assumption that has been made for the sake of simplicity is that the expected cluster size is the same for both types of samples. In general, the PPES sample will have larger clusters, but in the case where multi-stage sampling is used, we can assume the subsampling fractions are such that the number of second stage units per cluster in the respective samples is the same. The value of k may vary greatly from survey to survey. For example, suppose one is interested in new single-family residential construction. The cluster might be a permit office and an element would be the

individual permits issued over the previous year. If the variable of interest were the average number of square feet in a unit, this is usually available on the permit and so by sending an interviewer to sampled offices (first stage cost) and then enumerating the permits in that office (second stage costs), the value of k would be small. If, however, the variable of interest were not available on the permit, let us say the type of financing involved with the unit, it might be necessary to contact the owners of each unit or draw a subsample of units, thereby making the value of k quite substantial. In the tables below, values of 0, 1, 3, and ∞ are used for k to get an idea of the relative costs.

In Table 1, the values for "r" were empirically calculated from the four populations generated for this study. Since "r" is a function of the measures of size of the two subdomains, and is independent of the order of the units in the population, it does not affect the smoothed CV's obtained from each population.

Table 1 shows solutions to the equation above for differing values of k and "r" corresponding to ρ_0 = .96, .76, .54 and -.06, for n = 30. ($n_0 + n_1$) is the number of clusters surveyed for each domain and $n_0 + 2n_1$, for the entire survey.

In other words, consider the entries for k=1 and ρ_0 = .76, i.e., 27.24 (21.72). For the same cost as a single sample with 30 clusters, one would select an expected 21.72 clusters for which both types of elements would be enumerated, 5.52 (expected value) clusters in which only type 1 elements were enumerated and 5.52 in which only type 2 elements were enumerated. For any other value of n[3/], say n', we have $n_0' = \frac{n'}{30} n_0'$ and $n_1' = r n_0'$.

In Tables 2a and 2b the CV's corresponding to the sample sizes ($n_0 + n_1$) in Table 1 are presented for each domain in each of the four populations.

For a given sample size, the case k = ∞ gives the CV's for sampling each subdomain without regard to the other. Hence comparisons to the ratio estimates are for the same sample size, or equivalently, for sampling each subdomain individually.

The tables show that when PPES is a preferred estimate for k = ∞, it still seems to be a strong competitor for two variables when k = 0 and ρ_0 = .54. Even, in the worst case studied, where ρ_0 = -.06, the associated CV's are not very far from those for ratio estimation. For example, for Table 2a; column U2-Y1 for k = ∞, (i.e., 30 clusters) a single PPES sample would give a CV of 2.68, 18 percent lower than the ratio estimate. When the number of sampled clusters is reduced because of costs for a second variable (with k=1) and ρ_0 = -.06, the computed CV is 3.04, still 7 percent below the ratio estimate.

IV. Summary and Further Research

It should be emphasized that this study was not at all a rigorous treatment of the problem, but only an indication of whether this methodology should be explored any further. In practice one would generally encounter sampling units where the measures of size for the two types of elements (or variables) were reasonably well correlated. That is, the selection probabilities would not differ by a great amount. Judging from Tables 2a and 2b, the increase (from k = ∞) in sampling error for populations with only moderate correlation of the measures (.54 and .76) is in the range of 15-30 percent where second stage costs are negligible relative to first stage costs (i.e., k=0). If, then, second stage costs were not negligible (say k > 2) and PPES sampling gives better results than ratio estimates (i.e., 10 percent or more) for k = ∞, the dual sample approach 14 seems to give results that are still competitive with the single sample ratio estimation technique. Further research is needed to establish these results more rigorously so that one might use them with some degree of confidence.

It might involve comparing a closed form of the PPES variances to the ratio estimator, either for a specific order or over all possible orders, and also relaxing the assumptions made in the cost functions. Other avenues of research that may prove useful are comparisons with other competitive procedures, such as a single PPES sample proportional to a weighted average of the two measures of size. Also, the effect of varying the sample sizes for the two variables should be examined. For example, if we were to use a combined measure of size $M_i = w_1 M_{1i} + w_2 M_{2i}$ ($w_1 + w_2 = 1$) where the w's are the relative subjective importance of the variables, we might consider the dual approach where $n_1/n_2 = w_1/w_2$. The generalization of the procedure to more than two variables could prove fruitful.

In this study, systematic sampling was used throughout; however, it certainly was not necessary for the second variable. Other types of PPES without replacement sampling might also be examined which may further reduce the sampling error. One drawback of systematic samples is the problem of estimating unbiased variances. In respect to the first variable, systematic sampling was used because it severely restricted the distribution of the samples. Perhaps other PPES schemes might accomplish the same objective. As far as this aspect, that is the computing of conditional probabilities is concerned, one could test the limits of larger computers to determine how big a universe-sample size combination can reasonably be accommodated by the proce dure outlined in Section II.

FOOTNOTES

1/ The fixed probabilities, p_{1i} and p_{2i} impose the condition:

$$p_{1i}\, a_i + (1 - p_{1i})\, b_i = p_{2i}$$

or, $p_{1i}\, a_i = p_{2i} - (1-p_{1i})\, b_i$

under the constraints $0 \leqslant a_i \leqslant 1$; $0 \leqslant b_i \leqslant 1$

where a_i is the conditional probability of selecting unit i for sample 2, given it is

selected in sample 1, and b_i, given it is not selected for sample 1; $p_{1i} \cdot a_i$ is the probability that unit i is in both samples. The maximum value of $p_{1i} \cdot a_i$ is therefore achieved when $a_i = p_{1i}/p_{2i}$ $(b_i=0)$ for $p_{2i} < p_{1i}$ and $a_i = 1$ $(b_i = (p_{2i} - p_{1i}) / (1 - p_{1i}))$ for $p_{2i} \geq p_{1i}$.

2/ If the matrix does not converge, the problem can be solved as a transportation problem by dropping conditions (3) and (4). See Causey, et al.

3/ In this study $M_{1i}/\sum M_{1i}$ and $M_{2i}/\sum M_{2i}$ were less than 1 for all i. The above statement assumes n is small enough so that all units have selection probabilities less than 1.

4/ For $k = \infty$ there are no first stage costs, hence n units are selected for each variable (see Table 1) and the CV's of Y_1 and Y_2 are not a function of ρ_o.

REFERENCES

Cochran, W. G., (1973), "Sampling Techniques:" Wiley

Causey, B., Cox, L. and Ernst, L, (1983), "Application of Transportation Theory to Statistical Problems": Proceedings of the ASA, 1983.

Fagan, J. and Greenberg, B., (1984), "Making Tables Additive in the Presence of Zeros": to appear in Proceedings of the ASA, 1984.

Hartley, H. O., (1966), "Systematic Sampling With Unequal Probability and Without Replacement": JASA, Vol. 61, p. 740.

Table 1: Equivalent Expected Sample Sizes $(n_o + n_1)$ for n = 30

ρ_o:	.96 (r=.095)	.76 (r=.254)	.54 (r=.400)	-.06 (r=.637)
k				
0	27.81(25.62)	24.95(19.90)	23.31(16.62)	21.60(13.20)
1	28.86(26.59)	27.24(21.72)	26.26(18.78)	25.11(15.34)
3	29.53(27.21)	28.56(22.77)	28.00(20.00)	27.33(16.70)
	30 (27.64)	30 (23.92)	30 (21.50)	30 (18.33)

Values of n_o (the expected number of overlap units) are given in parenthesis.

Table 2a: CV's (%) for Equivalent PPES Samples n = 30

	U1		U2		U3		U4	
	Y1	Y2	Y1	Y2	Y1	Y2	Y1	Y2
ρ_i(i=1,2)	.987	.989	.951	.957	.839	.850	.687	.745
Estimates								
SRS Unbiased	10.23	12.76	10.52	11.13	11.67	12.28	14.09	16.53
Ratio	1.69	1.86	3.26	3.25	6.37	6.48	10.25	11.03
PPES:								
k ρ_o								
∞ 4/	1.66	2.03	2.68	3.04	5.91	6.48	8.87	9.40
3 .96	1.69	2.05	2.70	3.10	5.95	6.55	8.93	9.46
.76	1.74	2.09	2.77	3.22	6.04	6.74	9.14	9.60
.54	1.78	2.13	2.80	3.29	6.14	6.92	9.24	9.70
-.06	1.82	2.19	2.87	3.37	6.24	7.33	9.41	9.88
1 .96	1.71	2.10	2.72	3.18	6.03	6.72	9.09	9.56
.76	1.80	2.18	2.85	3.28	6.23	7.08	9.39	9.87
.54	1.87	2.23	2.95	3.50	6.40	7.32	9.60	10.07
-.06	1.94	2.30	3.04	3.65	6.57	7.53	9.88	10.30
0 .96	1.79	2.14	2.82	3.33	6.17	6.97	9.28	9.78
.76	1.95	2.33	3.07	3.69	6.62	7.62	9.92	10.32
.54	2.06	2.44	3.24	3.92	6.90	8.01	10.31	10.63
-.06	2.18	2.58	3.42	4.17	7.32	8.54	10.69	11.04

Table 2b: CV's (%) for Equivalent PPES Samples n = 50

	U1		U2		U3		U4	
	Y1	Y2	Y1	Y2	Y1	Y2	Y1	Y2
ρ_i(i=1,2)	.987	.989	.951	.957	.839	.850	.687	.745
Estimates								
SRS Unbiased	7.68	9.58	7.90	8.36	8.76	9.22	10.58	12.42
Ratio	1.27	1.39	2.44	2.44	4.78	4.87	7.70	8.28
PPES:								
k ρ_o								
∞	1.02	1.31	1.81	1.82	3.91	4.13	6.00	7.56
0 .96	1.08	1.38	1.83	1.83	4.18	4.36	6.40	7.73
.76	1.19	1.53	1.97	2.14	4.66	4.75	7.00	8.06
.54	1.29	1.62	2.12	2.33	4.92	5.03	7.38	8.30
-.06	1.39	1.72	2.27	2.50	5.21	5.41	7.82	8.60