METHODS TO CONTROL SELECTION WHEN SAMPLING FROM MULTIPLE LIST FRAMES

John Amrhein USDA-NASS, Susan Hicks USDA-NASS, Phil Kott USDA-NASS
John Amrhein, National Agricultural Statistics Service, 4818 South Bldg., Washington, D.C. 20250

KEY WORDS: Multivariate stratification, Permanent Random Numbers, Poisson Sampling, Collocated Sampling, Probability Proportional to Size Sampling

Introduction

The National Agricultural Statistics Service (NASS) conducts a variety of agricultural surveys using a prioritized stratification scheme for multiple commodities. Bankier (1986), Skinner (1991) and Skinner et al. (1994) showed how a method of combining independently drawn stratified simple random samples could be made more efficient than previously thought. This paper explores the use of methods that select similar samples across all stratifications, thus reducing the total realized sample sizes. Data from one survey, the Vegetable Chemical Use Survey, is used to evaluate the effectiveness of alternative sampling strategies.

NASS' traditional stratification strategy for multivariate surveys prioritizes the items of interest such that population units with the rarest items are grouped together first, then units with the next rarest item and so forth until the smallest units with only the most common item are placed in the last stratum. This works well enough for surveys in which the items of interest are few in number. However, the Vegetable Chemical Use Survey (VCUS) collects data concerning as many as 25 agricultural commodities. Stratifying such a population while maintaining control over commodity-specific sample sizes presents a rather interesting problem.

One strategy that has sparked our interest is presented by Bankier (1986) and involves stratifying for each commodity in its own frame, drawing independent samples and estimating across all frames. We will refer to this as the independent frame technique. This strategy will ensure commodity-specific sample sizes at least as large as is desired since a commodity's sample size cannot be smaller than that drawn from its own frame. However, it would be a desirable characteristic to have samples from frames include units that possess multiple commodities of interest so that fewer interviews would be needed. That is, the samples from the various frames would include the same units whenever possible.

One way to draw samples that increases the likelihood of drawing similar units across multiple frames is to coordinate the samples using permanent random number (PRN) methods. The remainder of this paper discusses the use of three PRN methods and a probability proportional to size (PPS) method towards this end. The four methods are briefly described and then applied to the sampling population for the VCUS in several states. The resulting realized sample sizes are then compared and discussed. Finally, some comments concerning probabilities of selection are made.

Three Alternative Permanent Random Number Techniques

Fixed Sample Size Technique Suppose we have F commodities of interest and therefore F sampling frames where each frame stratifies the population according to a different commodity. Each frame f has H_f strata. Sampling rates are determined so that each stratum in each frame has an n_{fh} desired sample size to be drawn from a population of N_{fh} units. Each unit is assigned to one stratum in each frame. It is convenient to think of each frame having a "zero" stratum containing all units in the population that do not possess the commodity for which that frame is stratified. No sample need be drawn from this stratum.

Each population unit is assigned a PRN drawn from the uniform distribution on the interval [0,1). Each unit keeps its same PRN across all frames. Then, within each stratum in each frame the units are sorted according to ascending (or descending) PRN. The first n_{fh} units in stratum h of frame f are chosen to be included in the sample.

The idea is that if a unit's PRN is "small" in one frame, it is likely to be "small" in other frames and so that unit is more likely to be selected from more than one frame than if samples were drawn independently across frames.

Poisson Technique The assignment of PRNs and the construction of frames is the same as that described above for the fixed sample size method. However, instead of a desired sample size, n_{fh} , being assigned to each stratum, a desired sample proportion, $p_{fh} = n_{fh} / N_{fh}$, is assigned. Then each unit in stratum h of frame f has its PRN compared to p_{fh} and if that unit's PRN is less than p_{fh}, it is selected to be in the sample. It should be noted that the realized sample size becomes random under this technique.

Collocated Poisson Technique A variation of the Poisson method is collocated PRN sampling (Ohlsson, 1995). This technique is designed to reduce the variability of sample sizes that are possible with the Poisson method. After the assignment of PRNs, each unit is given a rank based on the size of its PRN. The unit with the smallest PRN receives rank 1, the unit with the next larger PRN is rank 2 and so forth until all N units are assigned a rank. Then each unit is assigned another number, say R_i, such that R_i = (Rank of unit i - ϵ)/N where ϵ is a random number from the uniform distribution on the interval [0,1). Each unit then has its R_i compared to p_{fh} and is selected if R_i is less than p_{fh}. Thus, the sampling is the same procedure as for the Poisson method with R_i replacing the PRN. This adjustment of the Poisson method serves to evenly spread the population units to remove any clusters of PRNs. However, because the units are stratified in each frame after collocation, clustering of units may reappear.

A Systematic Probability Proportional to Size Design

The fourth method we evaluated is a PPS design. In this technique the "zero" stratum discussed earlier is not a matter of convenience but is required. Selection proceeds as follows:

1) Divide the population into mutually exclusive cells such that every unit in a particular cell is in the same stratum in each frame.
2) For each unit i calculate π_i, the largest value across all frames of n_{fh}/N_{fh} where i is in stratum h of frame f. (Note that π_i is constant within a cell).
3) Randomly sort the units within each cell and then sort the cells in any order.
4) Draw a systematic PPS sample from the list using 1 as the sampling interval and assigning each unit i the measure of "size" π_i.

The advantage of this design is that the realized sample sizes for a particular cell will be equal to the integer on either side of the sum of the π_is within that cell.

Evaluation of the alternative techniques

To evaluate these techniques, we selected three states that conduct the Vegetable Chemical Use Survey and replicated the three PRN techniques, the PPS method and the independent frame technique 100 times. The assigned PRNs were maintained across the three PRN techniques within each replicate. A separate frame was constructed for each commodity of interest within a state. Population units were allocated to one of four strata in each frame; two probability strata, one take-all stratum and one zero stratum were used in each frame. Strata boundaries were determined using a modified Lavallee and Hidiroglou method and units were assigned to strata based on a $cum^3\sqrt{f(x)}$ rule (Sweet and Sigman, 1995). This stratification was chosen to mimic what might be a reasonable or reasonably common univariate sample design.

A target sample size of one-third the population was selected from each of the probability strata. Table 1 compares the overall sample sizes realized from each of the sampling techniques. As expected, the independent frame approach realized the largest sample sizes. The three PRN techniques realized sample sizes of similar size with the Poisson method experiencing the highest standard deviations in each of 3 trials (states). The PPS method appears to be the most stable.

Table 2 shows the percentage of strata-level Poisson and PPS samples that fell short of their target sample sizes. One reason more shortfalls were not observed in the Poisson methods' realized sample sizes is the occurrence of what we call "visitors". A visitor is a sample unit that

State	Independent Frame Method	Fixed Sample Size Method	Poisson PRN Method	Collocated PRN Method	Systematic PPS Method
CA	496 (8.8)	388 (9.6)	375 (11.1)	374 (5.6)	373 (.14)
MI	658 (9.3)	513 (9.2)	504 (13.6)	501 (6.0)	502 (.48)
NJ	563 (8.1)	359 (8.6)	343 (13.8)	344 (4.6)	343 (.17)

Table 1. Mean realized sample sizes over 100 replications of sampling. Standard deviations are in parentheses. Population sizes are: CA-775; MI-1041; NJ-785.

State	Poisson PRN Method	Collocated PRN Method	Systematic PPS Method
CA	11 %	11%	6.3%
MI	12 %	12%	6.3%
NJ	11 %	8%	1.4%

Table 2. Percentage of strata-level realized sample sizes (in probability strata) that fell short of the desired sample size over 100 replications of sampling.

was not chosen within a specific commodity's frame, but ends up in the sample because it was selected in another commodity's frame. For example, a farmer might grow both carrots and tomatoes. For carrots, this unit's PRN might be low enough that it is selected in the carrot frame. However, because tomatoes are a common commodity, there are enough other units with smaller PRNs to meet the sample size requirements that the unit is not selected from the tomato frame. This unit was selected for carrots, but not tomatoes although it has both. This unit is then a "visitor" in the tomato frame. The unit will have its tomato data collected and used for estimation even though it was not selected for that commodity. Visitors often make up the difference for the variability of the sample size in the PRN sampling techniques.

Figure 1 shows cumulative distributions of differences between realized and desired sample sizes as percents of the desired sample sizes for the sampled strata. That is, the cumulative distribution of (realized - desired)/desired at the probability stratum level. For example, Michigan had 13 commodity frames each with two probability strata. Sampling from these frames was replicated 100 times so that the cdf for each technique utilized 2600 points. The two Poisson methods are shown as a single line since they coincide. The Poisson methods do not over-sample as much as the fixed sample size and independent frame methods, but at the risk of under-sampling as we saw in table 2. The fixed sample size techniques (with dependent and independent frames) do not experience under-sampling, but do experience more over-sampling than the Poisson and PPS methods. The PPS method experiences some under-sampling but not to the extent of the Poisson methods. The PPS design also shows the steepest gradient of all the CDFs, indicating that it realizes less over-sampling.

Probabilities of Selection

Let p_{fh} be the probability of selection for all units in stratum h of frame f. Let π_i be unit i's selection probability across all frames. Under the independent frame approach, a unit will have a different probability of selection, $p_{fh} = n_{fh}/N_{fh}$, for each frame. Under the Poisson and collocated techniques, the probability of selection for unit i is $\pi_i = \max_f(p_{fh})$ where h corresponds to the stratum in which i belongs for frame f. The same probability of selection is used for the PPS technique which, as mentioned earlier, is used as the measure of size when selecting the sample. However, the probabilities of

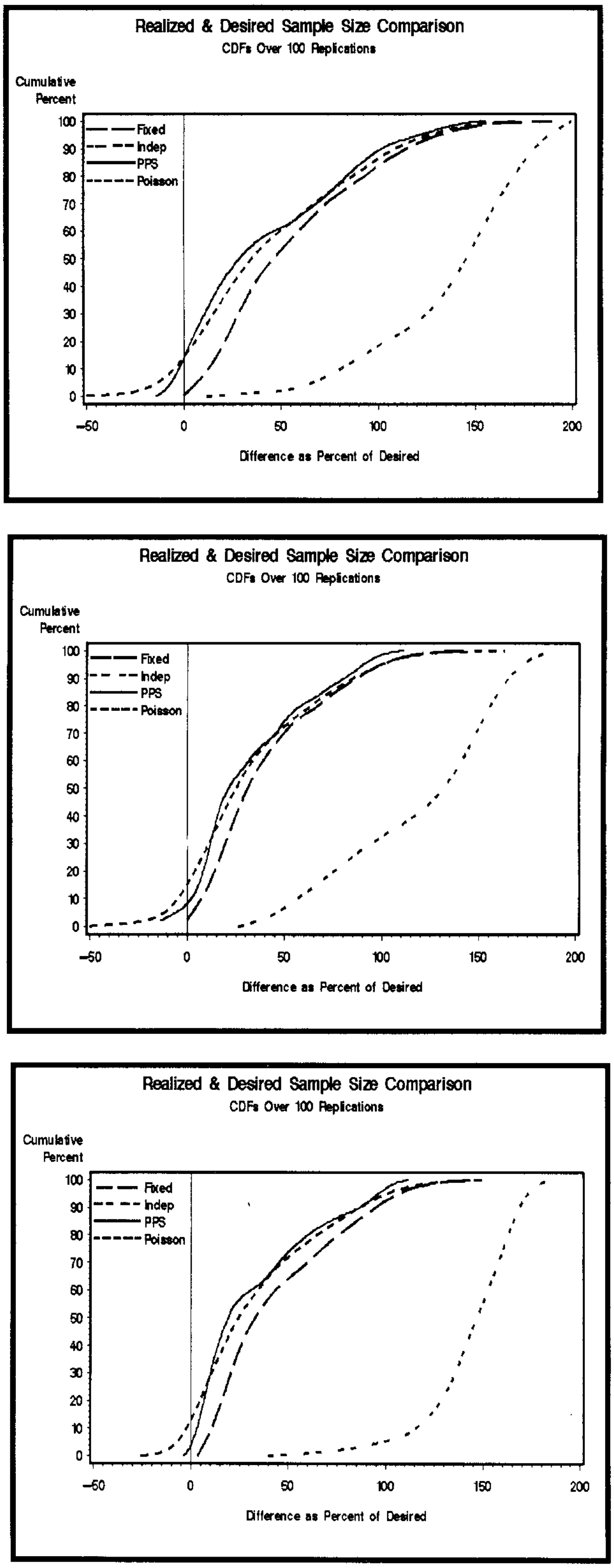

Figure 1 Comparison of realized and desired sample sizes for sampled strata. Top - CA; middle - MI; bottom - NJ.

selection under the fixed sample size PRN method are not obvious. It is possible for a unit to be selected from a frame with a lower probability than it has in another frame from which it was not selected. A unit's probability of selection considering all frames under the fixed sample size technique depends not only on its own assigned PRN but also on those of the other units in its strata.

We decided to simulate the probabilities of selection under this technique and compare the simulated results to $\pi_i = \max_f(p_{fh})$. Since all probability strata were sampled at a rate of 1/3, the simulated probabilities are compared to 1/3. The sampling technique was run 10,000 times using California's 1994 VCUS data. There were 19 commodities of interest in this state, but no units existed in probability strata in exactly 16 or 19 frames. The mean simulated probabilities of selection over the 10,000 trials are shown in figure 2 as a function of the number of frames to which units belong. A units probability of selection increases with the number of frames to which it belongs.

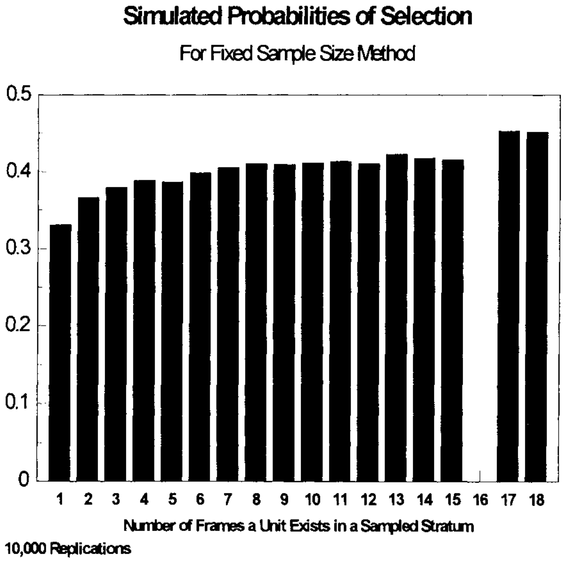

Figure 2 Simulated probabilities of selection for the fixed sample size method - California.

Estimation

A discussion concerning point and variance estimation under these sampling strategies is beyond the scope of this paper. The interested reader is referred to Kott (1996), in these proceedings, for a thorough treatment of estimation.

Conclusions

The systematic PPS design realized the most stable overall sample sizes in our simulations among the designs that experience variable sample sizes. The Poisson methods realized more variable sample sizes and more frequent under-sampling. The fixed sample size method does not realize any under-sampling, but has the disadvantage of requiring the determination of probabilities of selection via simulation. The independent frame method also meets all desired sample sizes but at the cost of grossly over-sampling.

Our plans for further research include testing the effects of various cell sorts (step 3 in the earlier discussion) on realized sample sizes in the PPS method. For example, first listing the cells pertaining to the rarest crop, then listing the cells for the second rarest and so forth.

References

Bankier, M. D. (1986), "Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys," *Journal of the American Statistical Association*, 81, 1074-9.

Kott, Phillip S. (1996), "Calibration Estimators Based on Several Separate Stratifications," *ASA Proceedings of the Section on Survey Research Methods.* (this volume).

Ohlsson, Esbjorn. (1995), "Coordination of Samples Using Permanent Random Numbers," in B. Cox et al. (eds.), *Business Survey Methods*, New York: John Wiley and Sons, Inc., p. 161.

Skinner, C. J. (1991), "On the efficiency of raking ratio estimation for multiple frame surveys," *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C. J., D. J. Holmes and D. Holt. (1994), "Multiple Frame Sampling for Multivariate Stratification," *International Statistical Review*, 62, 3, pp. 333-347.

Sweet, Elizabeth and Richard Sigman. (1995), "*User Guide for the Generalized* SAS *Univariate Stratification Program*," Economical Statistical Methods and Programming Division, Bureau of the Census, U.S. Department of Commerce, Report number ESM-9504.