ODDS RATIO SEQUENTIAL POISSON SAMPLING: A FIXED SAMPLE SIZE PPS APPROXIMATION

Pedro J. Saavedra, Michael T. Errecart and William Robb, Macro International
Pedro J. Saavedra, Macro International, 11785 Beltsville Drive, Calverton, MD 20795

Key words: Simulation, permanent random number, PPS sampling, variance

Permanent random numbers are used in survey sampling to control overlap between samples. While there is no exact PPS method for sampling using permanent random numbers with a fixed sample size, Sequential Poisson Sampling and Odds Ratio Sequential Poisson Sampling have been shown to provide adequate approximations. This paper reports simulations using national frames, and compares the efficiency and bias of the Goodman-Kish method, Poisson Sampling, Sequential Poisson Sampling and Odds-Ratio Sequential Poisson Sampling.

1. Introduction

When one is conducting more than one survey from similar or overlapping frames requiring different sample sizes or different stratifications, it is often desirable to control the overlap of the samples. For example, one might minimize the data collection effort by "piggy-backing" multiple questionnaires onto a single respondent. For rotating samples, it is often important to control the proportion of new units and the disposal of old units to reduce discontinuities in longitudinal estimates. Interviewers can be used more efficiently if many of the same primary sampling units are used across surveys. Equally important is negative coordination, ensuring that two sample draws do not overlap at all. Permanent random number (PRN) techniques can help with all of these problems.

Using PRNs is very simple when the design calls for simple random samples or stratified simple random samples. Ohlsson (1995a) offers a comprehensive discussion of this topic using permanent random numbers. To draw a simple random sample from a stratum or the entire population, one need only assign each unit a random number drawn from the uniform distribution between zero and one, treat the segment between zero and one as if it were a circle and select a starting point on the resulting circle. The elements in each stratum closest to the starting point in the positive direction are assigned to the sample until the sample size goal is met.

A variety of approaches have been suggested aimed at achieving a fixed-size sample drawn with probability proportional to size (PPS). In this paper we consider three techniques based on PRNs and, for comparison purposes, a non-PRN method. The methods are compared through simulation studies using two national sampling frames.

2. Sampling Techniques

Poisson Sampling

While the use of a permanent random number is not difficult when each element in a stratum has an equal probability of selection, the procedure is more difficult when one wishes to sample with probabilities proportional to size. One very easy approach to PPS sampling with a permanent random number is Poisson sampling. If n is the desired sample size and s_i the measure of size of unit i, we will define p_i as ns_i. A random number, r_i, from the uniform distribution is assigned to unit i, and the Poisson sample consists of those units for which $r_i \le p_i$. Poisson sampling has the advantage that the random numbers assigned to each unit can be treated as permanent and used to control overlap across different PPS samples. It has the disadvantage that the sample size can differ considerably from the desired sample size.

Sequential Poisson Sampling

Ohlsson (1990,1995b) suggested the Sequential Poisson sampling (SPS) procedure. SPS produces a fixed sample size, but yields probabilities of selection that are not exactly PPS. In SPS one calculates the quotient r_i/p_i, sorts by the quotients and selects the first n units.

Odds Ratio Sequential Poisson Sampling (ORSPS)

Saavedra (1995) developed a modification of Ohlsson's method based on the experience that odds ratios have often been found to be more useful than ratios of proportions. In ORSPS, the following quantity is used to sort the entries:

$$(r_i - r_i p_i)/(p_i - r_i p_i)$$

As for SPS, the first n entries are then selected.

At the same time the ORSPS approach was developed by Saavedra, Rosen (1996a, 1996b) was developing the procedure from a theoretical perspective. Rosen derived a family of sampling methods, and showed that the optimal procedure in the family was the equivalent to ORSPS, which he named Pareto Sampling. Rosen's work

came to the attention of Saavedra after the first draft of this paper had been submitted (the equivalence was pointed out by Ohlsson, 1996).

If r_i and p_i are such that a Poisson sample would yield a sample size equal to the sum of the probabilities, then Poisson, SPS and ORSPS yield the exact sane sample. In other words, when a Poisson sample yields a sample with an n equal to the sum of the probabilities of selection from the frame, the random numbers used to obtain that sample would yield the exact same sample if used with the SPS or the ORSPS method.

Goodman and Kish

The procedure of Goodman and Kish (G&K)(1950), described as Procedure 2 in Brewer and Hanif (1983), was also simulated for comparative purposes. G&K does not use permanent random numbers, but yields exact sample sizes at exact PPS probabilities. It suffers principally from the problem that there may be pairs of units with joint probability of zero. The G&K procedure works as follows.

Each unit in the frame is assigned a probability p_i such that the probabilities add up to the desired sample size, and the probabilities are all greater than zero and not greater than 1. The frame is then randomly ordered, and the i is taken to designate its place in the ordering. For each unit let c_i be the sum of the probabilities of the unit and all the preceding units. Select a random number x between 0 and 1. The sample is defined by the c_i which are such that $c_{i-1} < x+j < c_i$ for some $j = 0, 1, 2 \ldots n-1$.

3. Estimators

The probabilities of selection of units for SPS and ORSPS are not exactly PPS and there is no easy way of calculating them. Taking the p_i's as the nominal probabilities of selection, Ohlsson (1995a, 1995b) suggests the use of the following estimator:

$$(1/n) \sum_{i \epsilon S} (y_i/p_i)$$

Note that this is not the exact Horvitz-Thompson estimator since the probability of selection under this procedure has not been calculated. Ohlsson demonstrated through simulations that SPS is as efficient as Poisson sampling when this estimator is used. In addition, Ohlsson (1995b) showed that asymptotic normality holds for this estimator for SPS. This estimator was used for the simulations involving SPS and ORSPS.

The same estimator is, of course, the Horvitz-Thompson estimator for Poisson and Goodman-Kish. In addition, the Poisson estimator can be adjusted by n*/n where n* is the expected sample size and n is the actual sample size.

4. Method

Saavedra (1995) compared the four techniques in a simulation study involving sampling counties in two states. In that paper, he focused on comparing the empirical probabilities of selection of the counties to the true probabilities of selection under a PPS model. A Chi-square statistic and absolute differences were used to evaluate the results.

All four methods had chi-square that were not significant when one used the number of counties as degrees of freedom. At first some attention was paid to the fact that the Poisson method performed the best, but some additional simulations made it clear that the chi-square using cumulative simulations behaves like white noise, with first one method and then another having the lower chi-square. Other methods (e.g. ordering by p-r) yielded significant chi-squares.

This investigation focuses on comparisons of sample estimates to known population values, rather than focusing on the goodness of fit of actual selection probabilities to true PPS selection probabilities. It should be noted that it does not necessarily follow that the techniques that better reproduce the original probabilities will also yield lower mean square errors for sample estimates.

5. Simulation 1: Estimation of Petroleum Product Sales

The EIA-782B is a price and volume monthly survey which covers the fifty States and the District of Columbia, and includes sales of distillate, residual oil, motor gasoline and propane to end users and resellers. The EIA-782B does not include refiners, since these are covered by the EIA-782A (which includes all of them as a census). However, the refiners are included in the frame and are taken into account, since the sample design of the EIA-782B targets joint estimates of the EIA-782A and the EIA-782B.

The EIA-782B targets prices and volume CVs for residual oil, motor gasoline and propane for all fifty States (plus D.C.), but only targets distillate values for twenty-four States (and also for each of the Petroleum Allocation

Defense Districts--PADDs) . The frame for the EIA-782 is the EIA-863, a frame that includes volumes for a number of products.

We conducted simulations for the following seven products:

1) Residential Distillate
2) Nonresidential Retail Distillate
3) Wholesale Distillate
4) Retail Residual
5) Wholesale Residual
6) Retail Motor Gasoline
7) Total Propane

Since estimates are potentially desired for 51 States, five regions, three subregions and the nation, there are 60 geographic cells for which to derive estimates. As part of an approach to modifying the EIA-782 design, a set of probabilities of selection for each cell and product was obtained based on the proportion of the volume of the product a company sold in the cell, and the allocations for the cell in a previous survey. Through simulations using Sequential Poisson Sampling, the probabilities were modified through several iterations until a desired set of probabilities were established. The sample size was set at 2,000 and of these, 846 were certainties, either designated beforehand (e.g. refiners) or by virtue of their probabilities of selection.

Then 10,000 samples were drawn by each of the four methods. In addition, estimates were adjusted for sample size for the Poisson method. We focused on the US estimates, though we also recorded individual estimates for region and subregions. The efficiency of the simulations were measured by taking the absolute value of:

100 X (estimated - actual)/actual.

obtained for each of the seven products. Using ANOVA, and a Duncan test for difference between methods the following results were found. (Table 1 presents the Duncan tests).

For each product at least one of the Poisson sample estimates (adjusted or unadjusted) was worse than some of the other methods. There were no differences between SPS, ORSPS and G&K. Similar results were obtained when square differences were used instead of absolute differences.

In order to examine the possibilities of differences between SPS and ORSPS we also matched the samples which used the same PRNs. Again there were no differences. But it should be pointed out that a large proportion of the samples were identical, as Poisson would have yielded exactly the sum of the probabilities.

It should be pointed out that the reason why the Poisson estimator is not as good as the other is that Poisson estimates of totals should be made conditional on the sample size, but conditional Poisson sampling estimators have not yet been fully developed for PPS sampling, though work on these estimators is ongoing in Sweden (Ohlsson, 96b). The adjustments applied here may not be the most efficient.

The results suggest that either ORSPS or SPS should yield efficient results, at least for this type of frame, and selection of either should be based on their effect on the overlap of samples.

6. Simulation 2: The YRBS Study.

The next set of simulations were conducted using the frame for the Youth Risk Behavior Survey (YRBS), a survey sponsored by the Centers for Disease Control and Prevention. This is a national frame of secondary schools for which PSUs have been defined. The frame contains enrollment at each grade level (9th to 12th) and proportion of the student body which is Hispanic, African-American or Other. The simulations used the definition of PSU that were used in the YRBS Study, but in no way attempted to use the YRBS design.

In the YRBS frame each PSU had a minimum of three schools. The simulations sampled 600 schools, from 196 PSUs. PSUs were sampled with probabilities proportional to the number of schools in the PSU. Twelve PSUs were designated certainties and two were assigned more than three schools (one was assigned six and one twelve in order to compensate for their large size.

Four estimates were examined. One was the total student enrollment nationwide. The other three were the proportion of students nationwide in each ethnic group. To obtain the latter the estimates for each school were weighed by both sampling weight and enrollment.

This time there were no significant differences whatsoever. All three methods yielded similar results. Table 2 presents the results. All differences have been divided by the value obtained from the entire frame.

7. Conclusions

At this point the results are encouraging in so far as PPS sampling with a fixed sample size using permanent random numbers is concerned. In the simulations presented the approaches are as efficient as the Goodman and Kish method. However, unlike the G&K approach, the PRN methods permit rotation of the sample and control of overlap between samples. Hence, in cases where these issues become important, the use of SPS and ORSPS might be preferable. However, since the approach is asymptotic, more work needs to be done to identify situations where the sequential approaches might not work.

Ohlsson (1996b) reports that there are a number of cases where SPS and ORSPS (also known as Pareto sampling) yield similar efficiency, but there are other cases where ORSPS seems to yield better results. Given those findings, it seems preferable to use ORSPS in every case. But further empirical research is necessary.

Bibliography

Brewer, K.R.W., and Hanif M. (1983) Sampling with Unequal Probabilities, New York, Springer.

Goodman R. and Kish, L. (1950) "Controlled Selection - a Technique in Probability Sampling" J. Americ. Statist. Assoc. 45, 350-372.

Keyfitz, N. (1951) "Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities" J. Americ. Statist. Assoc. 46, pp.105-109.

Ohlsson, E. (1990) "Sequential Sampling from a Business Register and its application to the Swedish Consumer Price Index" R&D Report 1990:6 Stockholm, Statistics Sweden.

Ohlsson, E. (1995a) "Coordination of Samples Using Permanent Random Numbers" in Survey Methods for Business, Farms and Institutions, edited by Brenda Cox, New York: Wiley.

Ohlsson E. (1995b) Sequential Poisson Sampling Report No. 182, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, Stockholm, Sweden.

Ohlsson E. (1996a) Personal communication.

Ohlsson E. (1996b) Personal communication.

Rosen, B. (1995) On Sampling with Probability Proportional to Size. R&D Report 1995:1, Statistics Sweden.

Rosen, B. (1996) Asymptotic Theory for Order Sampling. R&D Report 1995:1, Statistics Sweden.

Saavedra, P.J. (1988) "Linking multiple stratifications: Two petroleum surveys." Proceedings of the 1988 Joint Statistical Meetings, American Statistical Association Survey Section, 777-781.

Saavedra, P.J. (1995) Fixed Sample Size PPS Approximations with a Permanent Random Number 1995 Joint Statistical Meetings, American Statistical Association, Orlando, Florida.

Acknowledgements

Thanks are given to Phil Kott for some helpful comments and to Benita O'Colmain and Michael Svilar for reviewing the manuscript.

Table 1: Petroleum Volume Simulations: Mean Absolute Relative Differences

PRODUCT	1. G-K	2. SPS	3.ORSPS	4.POISS.	5. APS*	DUNCAN
RESIDENTIAL DIST.	0.039210	0.039300	0.039285	0.043682	0.040149	4>2,3,1
NONRES. DIST.	0.028812	0.028587	0.028606	0.030538	0.028223	4>1,3,2,5
RESALE DIST.	0.012384	0.012319	0.012337	0.012845	0.016111	5>4>1,3,2
RETAIL RESIDUAL	0.011576	0.011646	0.011650	0.011775	0.016154	5>4,3,2,1
RESALE RESIDUAL	0.014964	0.014965	0.014961	0.014730	0.018325	5>2,1,3,4
RETAIL GASOLINE	0.013462	0.013245	0.013232	0.015625	0.014301	4>5>1,2,3
PROPANE	0.016235	0.016231	0.016230	0.017044	0.018068	5>4>1,2,3

* Adjusted Poisson Sampling estimate

Table 2: Youth Risk Survey Analysis: Mean Absolute Relative Differences

	1. G-K	2. SPS	3. ORSPS	DUNCAN
TOTAL	0.039415	0.038812	0.037542	no differences
AFRICAN-AMERICAN	0.079064	0.077904	0.080059	no differences
HISPANIC	0.114225	0.113972	0.117686	no differences
OTHER	0.021907	0.022085	0.021580	no differences