FIXED SAMPLE SIZE PPS APPROXIMATIONS WITH A PERMANENT RANDOM NUMBER

Pedro J. Saavedra, Macro International
Macro International, 11785 Beltsville Drive, Calverton, MD 20705

Key Words: simulations, working probabilities, overlap, Poisson sampling

Poisson sampling has the advantage that a permanent random number can be assigned to each unit, and used to link different PPS samples to control the overlap. It has the disadvantage that the sample size can vary considerably. Proposed approaches such as dividing a permanent random number by the probability of selection, sorting the frame using the quotient, and selecting the first n units do not yield exact PPS. Various approaches to obtaining a fixed sample size variant of Poisson sampling are considered, such as working probabilities, sorting by a function of both the sample size and the random number, and the use of a parameter to be adjusted so as to arrive at the desired sample size. These are evaluated through simulations by comparing empirically derived probabilities with the desired probabilities.

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. Perhaps one wishes to minimize the data collection effort by "piggybacking" questionnaires. Perhaps one is rotating samples, and wishes to achieve a certain overlap to avoid discontinuity. Perhaps one is conducting surveys from different populations and by defining the PSUs in the same way can use the same interviewers if the PSUs overlap sufficiently. Perhaps one wishes to insure a sample uncontaminated by a previous survey. One easy approach is the use of a permanent random number (PRN) assigned to each element in the frame.

There are other methods that do not use PRNs, such as the one devised by Keyfitz (1951). But these methods are much more restricted, and not as suitable for a long range situation involving a frame with births and deaths, whereas the use of PRNs is much more flexible.

Using a PRN is very simple when the design calls for a random or stratified sample. Ohlsson (1995a) offers a comprehensive discussion of the practice of sampling with stratified or random samples using a permanent random number. One merely assigns each element a random number between zero and one, treats the segment between zero and one as if it were a circle and selects a starting point. The elements in each stratum closest to the starting point in the positive direction are assigned to the sample until the allocations for the stratum are met.

The EIA-782 Petroleum Survey has used this strategy (Saavedra, 1988) with a twofold purpose. First, several different stratifications are obtained for the same population, corresponding to multiple estimates of interest. Then Neyman allocations are calculated for each stratification. Finally a PRN is assigned to each unit and a sample is drawn using this same PRN for each of the previously defined stratified samples. This guarantees that the different samples are drawn with a maximum overlap. Weights are obtained using simulations. Then, by changing the starting point the sample is rotated. Different approaches have been tried, but the most recent has been to calculate the maximum probability of selection of a unit across the different samples and to subtract a constant times that amount from each PRN, adding 1 to those which become negative.

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. There is, of course, one very easy approach to PPS sampling with a permanent random number, and that is Poisson sampling. One simply calculates the proportion of the total size represented by a unit (referred to as s), and multiplies it by the desired sample size n, and selects all those cases for which $r < sn$, where r is the PRN. Naturally, the approach could be used to sample by designating the desired probability p of each unit, but in the case of PPS sampling, $p=sn$. The only problem is that there is no guarantee that the sample size will be exactly n.

Brewer and Hanif (1983) present two modifications of Poisson sampling, but neither of them yields an exact fixed sample size (though a procedure called collocated sampling reduces the variance in sample size). There is no fixed sample size procedure which uses a random number such that the lower random numbers are more likely to be selected, and thus no procedure which may be used to link PPS samples using the same PRNs.

Ohlsson (1990,1995b) suggested sequential Poisson sampling (SPS) as an approach. This consisted of sorting the population by r/p ($=r/ns$) and selecting the first n cases. The probabilities of selection are not exact PPS, and there is no easy way of calculating them, but they constitute a good approximation.

The current investigation was motivated by the desire to improve on sequential Poisson sampling by finding an alternate method that would yield probabilities that are closer to being proportional to size. There seem to be at least three conceptually different approaches, of which only the third appears to be practical.

The first approach is to use working probabilities, in other words, to assign probabilities of selection which result in PPS probabilities when SPS is applied. If sequential Poisson sampling does not yield the desired probabilities, one could investigate what probabilities one would need to assign the elements to obtain the desired outcome using sequential Poisson sampling. This may be feasible for a very small population (say sampling two cases out of a population of four cases of normed sizes .1, .2, .3, and .4) but here the effort to be expended up front appears to be considerable.

A second approach is apparent if one defines Sequential Poisson Sampling as finding a t such that $\{x|r(x)<tp(x)\}$ has n elements, where r(x) is the PRN assigned to x and p(x) is the desired probability of selection (usually ns(x) where s(x) is the normed size of x). But tp(x) is not the only possible function of t and p(x) which could yield the desired result. For instance let the sampling design call for finding a t where $\{x|r(x)<f(t,p(x))\}$ has n elements and f is defined as :

$p(x)(t+1)$ for $t<0$
$p(x)$ for $t=0$, and
$p(x)+t(1-p(x))$ for $t>0$

Thus we can describe the second approach as finding an appropriate f which yields approximate or exact PPS. Yet finding such a function is not trivial, yet it is possible that the search for such a function would yield useful results.

The third approach is conceptually easier, as it involves using a different ratio in place of r/p. This approach has the advantage that it is easier to program and simulate, and different ratios can be investigated. One desirable characteristic would seem to be that if a Poisson sample should select n cases (the sum of the probabilities), the same n cases (using the same PRN) should be selected by the procedure. One way to accomplish this would be to apply the same transformation (monotonic between 0 and 1) to r and p. For example, one could sort by (1-p)/(1-r) which is applying $x'=1/(1-x)$ to both r and p. This, as it turns out, is not a very good candidate.

There still remains the criterion by which a method is to be evaluated. One approach would be to use simulations with some variable for which we seek an estimand. Using this approach Ohlsson found Sequential Poisson Sampling to yield a mean squared error slightly better than Poisson Sampling, using as weights the inverse of the theoretical probabilities.

This paper presents simulations at the first sampling stage, so a measure of the degree to which simulations using a particular method approximate PPS is needed. One good measure is based on the Karl Pearson Goodness of fit statistic (Hays, 1973). Over 10,000 simulations, the number of times each element is selected is observed, and the number expected given the desired probabilities of selection is calculated, then a chi-square statistic may be computed, (though the number of degrees of freedom is problematic when a method does not yield independence between the elements). At the very least the chi-square statistic can provide a descriptive measure of the fit of the method.

Ohlsson (1995b) suggested that the sum of the absolute deviations might be a better approach. The concern was that the chi-square method overemphasized convergence for the small units. It is true that a small discrepancy in probabilities for a small unit results in a large discrepancy in the unit weight. On the other hand, if the units are PSUs the PSU weight will be multiplied by the case weight, and the small discrepancy in weight for a large unit will potentially contribute more to the bias. Thus Ohlsson's suggestion was followed in addition to the chi-square method.

It should be explained that an inexact procedure could well outperform an exact one, since what these coefficients measure is the degree to which the procedure asymptotically approximates the desired probabilities. To see this consider two procedures for selecting approximately fifty cases out of 100 with equal probabilities. One is Poisson sampling, only the probabilities are slight altered in the fifth decimal place for a few cases. The second divides the 100 cases into two sets of fifty each and selects one or the other with probability of 1/2. It is easy to see that after 1000 samples, the proportion of the samples in which each unit was selected under the first method will be closer to .5 than for the second, because the units are selected independently. If the number of samples increases sufficiently, one would expect that the second approach would outperform the first, but this is of little or no practical significance.

Some preliminary simulations were conducted with small populations, but it was decided to use a real population in a problem which could conceivably arise. The selection of counties or PSUs in a state or national sample seemed an ideal one, and given that the ASA convention is taking place in Florida, the problem to be investigated was that of selecting five Florida counties (among the 67) with probabilities proportional to their 1990 census populations. The results thus obtained were then replicated by selecting six counties out of 24 Maryland counties also with probabilities proportional to their 1990 populations. The sample sizes were in part chosen so as to avoid certainty units.

Two approaches not involving PRNs with fixed sample size were used to place the results in perspective. The first was standard Poisson sampling, which uses PRNs but does not have a fixed sample size. The second was the random systematic procedure of Goodman and Kish (1950), described as Procedure 2 in Brewer and Hanif (1983), which does not use PRNs and suffers from the fact that there may be pairs of units with joint probability of zero, but which yields exact PPS probabilities. Sequential Poisson sampling (sorting by r/p) and several other procedures based on sorting by a function of r and p were also simulated.

Poisson sampling yielded the lowest chi-square (6.86) and the Goodman and Kish procedure yielded the third lowest (53.19). Sequential Poisson sampling was the fourth lowest (71.40) and three other methods yielded chi-squares in the thousands. But there was one method which performed better than SPS and Goodman and Kish, with a chi-square of 21.45 . The absolute difference coefficients yielded similar results.

The most effective fixed n procedure was obtained through a transformation of both r and p by the function $x'=x/(1-x)$. In other words, sorting by $(r/(1-r))/(p/(1-p))=r(1-p)/p(1-r)=(r-rp)/(p-rp)$. Table 1 presents the results of each set of 10,000 simulations. Since this procedure sorts by what would be the odds ratio if r were also a probability, the name Odds Ratio Sequential Poisson Sampling (ORSPS) suggests itself.

The four procedures with the lowest chi-squares were repeated selecting six out of the 24 Maryland counties with PPS, also using the 1990 census populations as a measure of size. The results (presented in Table 2, including county level results) were in the same order, only the performance of Sequential Poisson sampling was relatively worse than in Florida. Specifically Poisson sampling had a chi-square of 0.99, Goodman and Kish of 17.00, sequential Poisson sampling of 116.49 and the new procedure of 7.69.

It might seem strange that Goodman and Kish, being exact, is outperformed by ORSPS and the Poisson procedure. In fact the chi-squares for Goodman and Kish are approximately what one would expect. For some reason -- probably related to the independence of selection of units of the Poisson procedure -- the Poisson sample had a fit which was closer than what would have been expected.

Since this procedure seems to more effectively approximate the desired probabilities than the random systematic procedure of Goodman and Kish, one would surmise that the use of the estimator

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

where n is the fixed sample size and p the designated probability of selection should have a lower mean square error than the Goodman and Kish procedure. Note that this is not the exact Horvitz-Thompson estimator since the probability of selection by this procedure has not been calculated.

In conclusion, the modified Poisson Sequential sampling procedure appears to be very promising, and seems to provide a practical method of selecting a PPS sample with a PRN. This should provide a useful method for controlling overlap of samples, even where the frames merely overlap and different measures of size are used.

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.

Hays, W.L. (1973) Statistics for the Social Sciences, 2d Edition, New York: Holt, Rinehart and Winston.

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. (1995c) Personal communication.

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

Table 1
PPS Selection of five counties from Florida's 67 counties
10,000 Simulations for each method

	G-K	POISSON	SPS	ORSPS
Chi-square	53.19	6.86	71.40	21.45
Absolute difference*	99.58	19.58	145.67	55.96

* Multiplied by 1,000.

Table 2

PPS Sampling of six Maryland Counties out of 24
10,000 Simulations

County	Theor.	G-K	Poisson	SPS	ORSPS
24031	0.9500	0.9512	0.9508	0.9143	0.9551
24510	0.9236	0.9248	0.9239	0.9021	0.9231
24033	0.9151	0.9141	0.9156	0.8942	0.9139
24005	0.8685	0.8697	0.8688	0.8714	0.8715
24003	0.5361	0.5347	0.5359	0.6012	0.5428
24027	0.2351	0.2343	0.2350	0.2459	0.2347
24025	0.2285	0.2321	0.2284	0.2378	0.2291
24021	0.1885	0.1841	0.1885	0.1895	0.1850
24013	0.1548	0.1556	0.1548	0.1557	0.1561
24043	0.1523	0.1503	0.1520	0.1536	0.1493
24017	0.1269	0.1260	0.1270	0.1281	0.1267
24037	0.0953	0.0964	0.0952	0.0941	0.0939
24001	0.0940	0.0904	0.0941	0.0947	0.0951
24045	0.0933	0.0942	0.0948	0.0920	0.0938
24015	0.0895	0.0885	0.0876	0.0859	0.0865
24009	0.0645	0.0648	0.0638	0.0627	0.0633
24047	0.0440	0.0392	0.0440	0.0441	0.0447
24035	0.0426	0.0411	0.0421	0.0421	0.0421
24041	0.0383	0.0392	0.0381	0.0377	0.0369
24019	0.0379	0.0411	0.0384	0.0363	0.0349
24023	0.0353	0.0368	0.0352	0.0349	0.0356
24011	0.0339	0.0363	0.0337	0.0324	0.0347
24039	0.0294	0.0307	0.0297	0.0282	0.0288
24029	0.0224	0.0244	0.0220	0.0211	0.0224
Chi-square		17.00	0.99	116.49	7.69
Absolute dif		429.78	93.3	1863.5	400.66