IMPLICIT STRATIFICATION AND SAMPLE ROTATION USING PERMANENT RANDOM NUMBERS

Pedro J. Saavedra, Macro International Inc. and Paula Weir, Energy Information Administration
Pedro J. Saavedra, Macro International, 11785 Beltsville Dr., Calverton MD 20705

Key Words: Poisson sampling; Rotation; Geographical strata; PPS sampling

There is a class of sampling techniques which can be collectively described as order sampling. In this class of techniques a random number is assigned to each member of the frame and the sample is in some way drawn so that among units with similar characteristics (stratum, size, etc.) those with the smallest numbers will be selected. When the random number is preserved in order to control the overlap of the sample with a second sample from an overlapping frame, we speak of a Permanent Random Number (PRN). While the techniques presented in this paper were motivated by the use of Permanent Random Numbers, the first of these techniques is relevant to any form of order sampling, whether or not the numbers are used to preserve the overlap with other surveys. However, since there are better ways to achieve the objective than using order sampling when PRNs are not needed, the technique is particularly relevant to order sampling.

The second technique is relevant when Permanent Random Numbers are used to control overlap and one wishes a particular kind of overlap (say close to 50%) at all levels of the sample (i.e., regardless of size or stratum). Both of these techniques were used in the new design of the EIA-782 and are especially useful in an entire class of sampling designs.

Order Sampling

By *order sampling* we mean any procedure which can be accomplished through the use of a random variable with the frame as domain and (usually) the open interval (0,1) as the range, where, other things being equal, the lower values of the number are selected for the sample. The most common form of order sampling is *simple random sampling*, where a random number is assigned to every unit in the frame and the units with the n lowest numbers are selected for the sample.

Order sampling can be applied to *stratified sampling* where the n_j units with lowest numbers in stratum j are selected for the sample. *Poisson sampling* is another form of order sampling, where the sample is drawn with equal or unequal probabilities. In the case of equal probability, where the random number is between 0 and 1, the sample is drawn by selecting all units whose random number is lower than a fixed value. For an unequal probability sample different probabilities (proportional to some measure of size in most cases) are assigned to each unit. The unit whose random number is lower than its probability of selection is included in the sample. The main drawback of Poisson sampling is that it yields a variable sample size. The sum of the probabilities yields an expected sample size, but the size itself can vary considerably from its expectation.

There is a variant of Poisson sampling known as *collocated sampling* (Brewer and Hanif, 1983). In this case the random numbers are first converted to ranks, and then the number (R-.5)/N, where R is the rank and N the number of cases in the frame, is treated as the random number is treated in Poisson sampling. This has the effect of assuring that the (0,1) interval is divided into N equal segments and the random numbers used reflect the midpoint of those segments. This has the effect of reducing the variation in sample size. An alternative to this approach is to subtract a random number between zero and 1 from the rank. We will examine the usefulness of this variant in a subsequent discussion.

There are two more instances of order sampling that are of particular interest as they represent the former design and the current design of the *EIA-782 Petroleum Product Survey* (Saavedra, 1988; Saavedra and Weir, 1997). The first method (referred to in the survey as *linked sampling*) is useful when one has a multipurpose survey and multiple stratifications. Assume that a survey is designed to obtain estimates for several variables, and one has a value for a related variable (perhaps a previous year value) for each. Consider a separate stratification and separate Neyman allocations for each variable to be estimated. A single PRN is chosen to select a sample for each stratification. A unit is selected if it is chosen for any of the stratified samples. Unfortunately, there is no obvious analytic method for the calculation of the probabilities of selection, so this approach requires the use of simulations to estimate the probability of selection of each unit, and thus obtain Horwitz-Thompson type estimator weights for all units. This was used for many years as the method used for the EIA-782, and it has also been explored by the Department of Agriculture as well.

The second approach is really a series of approaches that approximate Poisson sampling, but yield a fixed sample size design. Ohlsson (1995) developed *sequential*

Poisson sampling where the noncertainty units in the frame are sorted in ascending order by r/p, where r is a random number associated with the unit and p is a probability of selection. If n is the sum of the noncertainty probabilities in the frame, selecting the first n cases is an approximation to Poisson sampling (and yields the same exact result if it turns out the Poisson sample would yield n cases).

Rosen (1995) and Saavedra (1995) discovered independently a refinement of Ohlsson's method which Rosen called *Pareto sampling.* In Pareto sampling the noncertainty units are sampled using the formula:

$$(r-pr)/(p-pr).$$

Rosen demonstrated that this is an optimal order sampling method with unequal probabilities and fixed sample size.

The EIA-782 currently uses Pareto sampling. However, the examples presented in this paper make use of the probabilities of selection and geographic cells of the EIA-782, but use Poisson sampling instead of Pareto sampling.

The EIA-782 Survey

The EIA-782 Monthly Petroleum Product Sales Report collects State level prices and volumes of petroleum products by sales type from all refiners and a sample of resellers and retailers. The data collected are aggregated to produce approximately 30,000 estimates and are published in the Petroleum Marketing Monthly. For each of ten targeted product/end-use categories, the noncertainty group was stratified by sales volume and urbanicity and then sampled within each stratum. A select set of State level average prices was targeted at a 1% Coefficient of Variation (CV) for determining sample sizes. These price CVs roughly correspond to volume CVs of 10% or 15%, depending on the petroleum product.

In the previous design of the EIA-782 Neyman allocation was used to determine the sample size required for each targeted product/end-use category. A triennial survey of all sellers of petroleum products provided State level sales volumes at the targeted levels and was used as the sampling frame and basis for stratification. Sample selection was carried out using a linked sample selection. In this process a respondent was selected randomly from the frame and used simultaneously to satisfy the required allocation in each of the targeted products. If the respondent's stratum had already reached the required allocation for one or more, but not all, target variables, the respondent was considered to be a volunteer or visitor for those variables. In the target variables for which the respondent helps to satisfy the allocations, the respondent was considered to be in the basic sample. If the respondent was not selected for a given State, he was considered a visitor for all stratifications in that State. The linked selection reduced the overall sample size by using each selected respondent to satisfy multiple requirements. Because the selection was not independent, the probability of selection for a sampled unit could not be calculated directly. Instead, the probabilities were derived by simulating 1,000 sample selections and counting the number of times each respondent was selected. The inverse of the frequency of selection divided by the number of simulations was used as the sample weight for estimation.

The new EIA-782 design involves designation of refiners and companies selling a high proportion of the volume of any target product in a State or region as a certainty unit. Then a probability of selection is assigned to each company for each product and publication cell where the product is sold. The calculation of each company's probabilities of selection for each of the 600 potential cells is an iterative process. Initial allocations were set at the previous sample's allocation. If a cell was not designated a publication cell, an allocation of zero was used. Given those allocations, for each company and cell, the company's volume is converted to a proportion of the total volume for that cell and multiplied by the initial allocation to obtain the probability of selection. The initial total sample size is then examined. If the size is too large or too small, the allocations were adjusted. This is done by preserving the certainty companies and multiplying the noncertainty companies by a constant. The initial probabilities were used in 100 simulations. Volumes were estimated for each cell from the 100 samples. The 100 trials were sufficient to obtain a clear picture of the percentage of an estimate. CVs were also calculated and examined. Allocations were then increased where CVs were too high, and decreased if CVs were unnecessarily low.

The examples presented in this paper are derived using actual data from the frame of the EIA-782, and the actual probabilities of selection and geographical strata are used. However, the examples will use Poisson sampling, since it is better known and easier to program for multiple simulations.

Implicit Stratification through Collocation

Every sampling statistician is familiar with implicit stratification when one uses the Goodman- Kish sampling approach. In this approach one selects a sampling interval, and if one wished to guarantee representation by some implicit strata which is proportional to the sum of the size measures of the strata, one simply groups the units by strata. For example if one were to sample school proportional to their enrollment and wanted the States represented in proportion to their total enrollment one simply sorts the schools by State and applies the sampling interval the same way as if the schools were sorted randomly or in some other order. Unfortunately this approach cannot be implemented using a PRN to control overlaps with subsequent surveys.

The equivalent approach for order sampling (and specifically for Poisson sampling and its variants) is to collocate separately for each of the implicit strata. When one collocates separately for subpopulations, one must subtract a random number rather than .5, since otherwise ties can easily result from different implicit strata.

Let us describe how this takes place. There were 26,264 noncertainty companies in the EIA-782 frame, with an expectation of 1,180 companies in the sample. There was a desire to achieve proportional representation by home State, and within home State by urban (MSA)/rural (non-MSA) status. There were 100 cells so defined. In order to do that, a random number was generated first. Within each cell, the companies were ranked using that random number. Then to each company the number r'=(R-s)/N was assigned to the company, where R was the rank of the company in its geographic cell, s a random number between 0 and 1, and N the number of companies in the geographic cell.

Another way of describing it is as follows. If there were N companies in a given cell (e.g., urban Texas) the (0,1) segment is divided into N equal segments. The company with the smallest r is assigned a random number within the first segment. The next company is assigned a random number within the next segment, and so forth.

Two hundred sets of random numbers were generated. From each set three samples were drawn. One was a standard Poisson sample. One was a sample using collocation at the State-urban status level and one using collocation at the national level. For each cell and nationally, three statistics were obtained. One was the standard deviation of the sample size within the cell, one was the absolution deviation of the sample size from the expected sample size (defined as the sum of the probabilities of the companies in the cell), and the third was the root mean square of the differences between the sample size and the expected sample size for the cell. The results were very similar regardless of what statistic one used, so only the third set of figures will be presented in Table 1.

A quick inspection of the results leads to obvious conclusions. At the national level cell specific collocation is about as effective as nationwide collocation. But at the cell level, the cell specific collocation is better in 99 out of 100 cells. The exception was the District of Columbia which has only a handful of noncertainty companies. Treating the cells as independent cases one gets significance at the .0001 level if one uses a matched pairs test

Implicit stratification through collocation was first used in the EIA-782 under the linked sample design. In that case cells were defined for Company-State units (CSUs) as "urban, rural and out-of-state". Later that stratification was made explicit.

Implicit stratification through collocation has several advantages:

1) It can be used for any kind of order sample.

2) It need not be applied to the entire population. One may choose to not classify or collocate certain units one has no information on. No bias exists for such a unit.

3) Cases may be added or subtracted from a cell, and the basic order preserved among the cases that remain or had been present, while including the new ones and preserving the proportional representation of the sample.

4) The method can be imposed over other existing stratifications or strategies.

Rotation of the Sample

Often a survey has a number of cycles, and it is desirable to achieve a certain overlap between consecutive cycles. This assures that one will not have a completely new sample, with a maximum of discontinuity between estimates. In a simple random sample it suffices to randomly select out half the sample and to replace it with an equal number of cases drawn randomly from the unsampled units. In a stratified sample one may do the same for each stratum separately. In an equal probability random sample with a Permanent Random Number, one would rotate the sample by subtracting a constant from each PRN and then adding 1 to the negative numbers. Thus r'=r-c and if r'<0 then r'=r'+1.

We examine in this section a procedure for use with an unequal probability Poisson sample, though the procedure may be used with Pareto sampling, and with a number of other unequal probability forms of order sampling. The first inclination would be to treat this case in much the same way as the equal probability case -- by subtracting a constant from every PRN, adding 1 to the negative numbers and thus obtaining a new PRN through which a new sample will be selected. The problem here is that such a strategy will inevitably rotate the small units out of the sample, while preserving the larger units. Some times this may be desired, but other times one may wish to rotate all kinds of units at a similar pace. The solution is to rotate by a product of the probability of selection. Thus r'=r-pc where p is the probability of selection and c is a constant, and where as before, if r' <0 then r'=r'+1.

Simulations were done with ten different coefficients, varying c from .05 to .5 for the constant wne r'=r-c was used, and from .1 to 1.0 when c was multiplied by the probability of selection. The frame was divided according to the probability of selection into classes going from p<.1 to p>.7and p< .8, with each interval of size equal to .1. The numbers examined were the proportions of the old sample present in the new. These results are presented in Table 2. It can be seen that in the first case all the overlap soon centers on the larger units. On the other hand, using the second method the overlap is uniformly present across size categories. This not only assures rotation at all levels, but it avoids the repeated appearance of noncertainty units cycle after cycle.

Bibliography

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

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

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

Saavedra, P. J. (1988) "Linking Multiple Stratifications: Two Petroleum Surveys". 1988 Joint Statistical Meetings, American Statistical Association, New Orleans, Louisiana.

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

Saavedra, P.J. and Weir, P. (1997) "The Use of a Variant of Poisson Sampling to Reduce Sample Size in a Multiple Product Price Survey", 1997 Joint Statistical Meetings, American Statistical Association, Annaheim, California.

Table 1: Sample Size Variations Using Three Methods

State	Urbanicity	Poisson	Implicit	Collocated
US	-----------	34.28	28.09	28.04
AK	RURAL	1.74	1.43	1.76
AK	URBAN	0.54	0.48	0.53
AL	RURAL	3.47	2.74	3.49
AL	URBAN	3.51	2.92	3.46
AR	RURAL	3.43	2.81	3.40
AR	URBAN	2.20	1.81	2.23
AZ	RURAL	1.24	1.14	1.21
AZ	URBAN	1.90	1.54	1.90
CA	RURAL	1.76	1.41	1.78
CA	URBAN	4.34	3.79	4.29
CO	RURAL	2.84	2.16	2.77
CO	URBAN	2.92	2.28	2.93
CT	RURAL	1.63	1.10	1.63
CT	URBAN	4.83	3.82	4.84
DC	URBAN	0.43	0.45	0.43
DE	RURAL	1.12	0.71	1.12
DE	URBAN	2.04	1.59	2.03
FL	RURAL	1.99	1.47	2.01
FL	URBAN	3.67	2.89	3.65
GA	RURAL	3.97	3.20	4.05
GA	URBAN	3.72	3.14	3.79
HI	RURAL	0.65	0.53	0.63
HI	URBAN	0.44	0.34	0.44
IA	RURAL	6.10	5.31	5.97
IA	URBAN	3.05	2.55	3.12

Table 2: Rotation by a Constant

CLASS	N	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ALL	26264	0.64	0.47	0.35	0.26	0.19	0.14	0.10	0.08	0.06	0.05
0-.1	19923	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.1-.2	2897	0.62	0.26	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.2-.3	1486	0.81	0.64	0.40	0.19	0.04	0.00	0.00	0.00	0.00	0.00
.3-.4	863	0.83	0.68	0.55	0.42	0.29	0.12	0.02	0.00	0.00	0.00
.4-.5	472	0.84	0.76	0.67	0.58	0.44	0.36	0.20	0.09	0.03	0.00
.5-.6	374	0.87	0.78	0.71	0.61	0.51	0.43	0.35	0.27	0.21	0.16
.6-.7	192	0.95	0.88	0.81	0.72	0.65	0.56	0.50	0.49	0.46	0.45
.7-.8	57	1.00	0.98	0.92	0.85	0.76	0.70	0.69	0.72	0.73	0.68

Table 3: Rotation Proportional to Probability of Selection

CLASS	N	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
ALL	26264	0.88	0.78	0.69	0.59	0.49	0.41	0.30	0.22	0.13	0.05
0-.1	19923	0.88	0.78	0.67	0.56	0.47	0.38	0.28	0.20	0.11	0.00
.1-.2	2897	0.89	0.78	0.67	0.56	0.45	0.36	0.24	0.15	0.06	0.00
.2-.3	1486	0.91	0.81	0.73	0.64	0.54	0.44	0.30	0.23	0.12	0.00
.3-.4	863	0.87	0.75	0.66	0.58	0.48	0.41	0.32	0.18	0.07	0.00
.4-.5	472	0.85	0.78	0.70	0.62	0.52	0.41	0.32	0.19	0.07	0.00
.5-.6	374	0.86	0.75	0.68	0.56	0.47	0.39	0.28	0.23	0.16	0.15
.6-.7	192	0.92	0.84	0.73	0.64	0.55	0.49	0.44	0.44	0.47	0.45
.7-.8	57	0.98	0.95	0.85	0.70	0.70	0.73	0.68	0.72	0.71	0.73