The use of Permanent Random Numbers in a Multi-Product Petroleum Sales Survey: Twenty Years of a Developing Design

Pedro J. Saavedra and Paula Weir
Pedro J. Saavedra, ORC Macro, 11785 Beltsville Dr., Calverton, MD 20705
Pedro.J.Saavedra@orcmacro.com - PAULA.WEIR@eia.doe.gov
Fax (301) 572-0984

Keywords: permanent random numbers, establishment, Poisson sampling, rotation

Abstract: The EIA 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. In 1983 the survey was redesigned by a team of statisticians. The problem of designing a survey to obtain target coefficients of variation (CV) for multiple products and geographical locations was solved by using multiple stratifications linked by a Permanent Random Number (PRN). The PRNs were also used to rotate the sample. Two subsequent developments, the discovery of Pareto Sampling and the Chromy Allocation Algorithm, contributed to a later redesign of the EIA-782 in the middle and late 1990s. A number of issues associated with selection using PRNs, implicit stratification and the sample rotation process have been examined over the years. This paper describes the evolution of a multi-product establishment survey based on the use of Permanent Random Numbers as a sampling tool.

Early Stages

Permanent Random Numbers have been used in survey sampling in the United States since the first half of the Twentieth Century (eg. Perlman and Mandel, 1944), but only in the last ten years has there been an increased interest in their use in sampling. Brewer discussed them in an early paper, but did not use the name (Brewer et. al. 1972). Ebjorn Ohlsson (1995) introduced the term in the 1990s. The work of Bengt Rosen (1995) Anders Holmberg (2001), and Nibia Aires (1999) in Sweden, and of Larry Ernst (1999) at the Bureau of Labor Statistics has been responsible for renewed interest in controlling sampling overlap through PRNs. Long before the latest developments, the team that was redesigning the Monthly Petroleum Product Sales Report survey made use of the concept of Permanent Random Numbers and this procedure has driventhe evolution of the design over the last twenty years.

Back in the early 1980s, a group of statisticians[1] were given the task of redesigning a sample for the petroleum product price and volume survey of retailers and resellers that required a given CV for different States, regions and products. For the first year of the survey, only one product had been collected from non-refiners. There was concern when the survey was expanded to include non-refiner coverage for other products, and that the existing design was not sufficiently robust, given the large number of estimates to be derived from the sample. It was clear that a new design was in order.

Estimates were now required for three petroleum products (distillate, gasoline and residual oil), two or three end-uses and up to fifty States, the District of Columbia and seven Petroleum Allocation Defense Districts (PADDs) for each product. While large multi-State companies and refiners could be designated certainties, that would leave many small companies selling one, two or three products in one, two or three States, with no apparent way of integrating them into one sample. A triennial census 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.

The first suggestion was to create separate stratified samples for each product, using Neyman allocations, and draw separate samples, treating the Company-State Unit (CSU) as the sampling unit. CVs of 15% for distillate and 10% for the other products were found to yield the desired variances for prices. Since there were two or three end-uses for each product, separate allocations were obtained for each end-use, and the maximum was allocated for each cell. However, selecting independent samples for each product required a fairly large total sample. At this point, a member of the team suggested that the samples be selected by randomly ordering the frame and filling out the stratified samples using the same random ordering. Since the ordering was to be done by assigning a random number to each CSU, the use of PRNs

had in fact been proposed without using the term. The design was referred to as a "linked samples" design (Weir, 1984; Saavedra, 1988).

The idea was accepted by the team, particularly when simulations showed that the total number of companies and CSUs involved could be greatly reduced over the alternative of independently sampling for the three different products. Unfortunately, a complication soon arose. In order to simplify the collection program, it was decided to use only one form and require every company sampled to respond to the entire survey. This created a problem, as the use of one random order for selection would result in some cells having reached the desired allocation while additional units were needed for other cells. The design initially called for only reports from companies sampled for a given product and State to be used in the estimates for that product and State. This proved unacceptable, and a way to use all the data collected became necessary.

The initial proposed solution treated each unit as a certainty unit for every State and product for which it had not been sampled, but it was soon shown to not be viable. It was instead suggested that a company's probability of selection be calculated and a Horvitz-Thompson type estimator be used. This felicitous suggestion won approval, until it was time to find a formula to calculate each company's probability of selection. This effort proved unsuccessful, so it was decided that simulations be used to obtain empirical probabilities of selection. In those days of slow mainframes, for one thousand simulations, ten programs had to be submitted drawing one hundred samples each. Each run combined ten runs of 100 simulations each and used about 80 CPS minutes. The absence of an analytic formula to calculate probabilities was one of the two major challenges of the sample design (variable sample size being the other).

Thus, sample selection was at first carried out using cross-stratified samples linked by a PRN (later each end use was separately stratified and linked). In this process, a respondent was selected randomly from the frame using the PRN 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 for those variables. Likewise, if the respondent was not selected for a given State, he was considered a volunteer for all stratifications in that State. In target variables for which the respondent helped satisfy the allocations, the respondent was considered to be in the basic sample for the particular product. Otherwise, the respondent was called a "volunteer". Variances were calculated using only the basic sample and the strata for the particular product. The use of a PRN reduced the overall sample size by using each selected respondent to satisfy multiple requirements.

The first sample selection cycle used the Horwitz-Thompson type estimators with 1,000 simulations to determine the selection probabilities. Later, an adjusted estimator was developed, so that the inverse of the selection probabilities were adjusted for each product such that the sum of the weights within a cell (in the stratification associated with the product) would equal the population of CSUs in the cell..

Rotation of the Sample

The next major design hurdle was the rotation of the sample. In order to preserve data continuity while eventually shifting the burden from one set of respondents to the next, it was necessary to rotate fifty percent of the non-certainty portion of the sample each sample cycle. That is, the ideal objective was to have 50% of the non-certainty portion of the sample in a given cycle be rotated out for the next cycle, while attempting to have respondents stay in the sample for two cycles. The usual approach – rotating within cells – did not apply in light of the multiple stratifications.

The rotation issue shifted the focus from the ordering of the frame to the actual random numbers used to order the frame. It was decided that a constant z was required, so that if x were a random number such that $0 < x < 1$, then $x'=x-z$ if $x>z$ and $x'=x-z+1$ if $x < z$. The value of z was selected so that if x' were used to draw the new cycle, there would be a fifty percent overlap between the cycles. This, of course, meant that the original PRNs would be transformed and, in fact, after each cycle a record of only the new cycle's and the two previous cycles' PRNs was kept. That made the "permanent" in Permanent Random Numbers a relative matter.

However, this rotation scheme proved unsatisfactory, because the rotation would take place at different rates for cells with different sampling fractions. This led to the decision to make the rotation dependent on the probability of selection. Defining q as the maximum probability across the three stratifications, a constant v was chosen so that x'=x-vq or x'=x-vq+1, whichever falls between 0 and 1. This refinement meant that large and small companies (except for certainties) would be rotated at the same rate.

Another problem was that a group of CSUs from the same stratum could have clustered PRNs so that some would not be rotated out. In order to address this problem, the design began to use collocation, thus insuring an even spread of the PRNs. During one rotation in particular, it was feared that clustering in one product's cells would interfere with rotation. The PRNs were then ordered by cells. If a cell had n units, each unit in the cell was assigned an interval of equal length, and then a random number within that interval. Thus the resulting transformed PRNs were in the same order as the original cells, but spread throughout the (0,1) interval.

Sample 2000

With a reduced budget in the middle 1990s, the focus shifted to reducing survey operations' costs. These costs were directly associated with the sample sizes because operations' efficiencies were said to have already been fully realized. It was determined that the expected budget in 1997 would be sufficient to operate a sample of approximately 2,000 companies, compared to the existing sample of approximately 3,000 companies. This sample was appropriately named Sample 2000. An operational sample 66% as large as the current one represented a tremendous decrease, and it was expected that some target variables would be dropped and CVs would be increased for other variables. Also, requirements for State level estimates for all States and variables were expected to be loosened to reach the reduced sample size. It became apparent that a new design was essential.

For some time the team had been seeking a way of selecting a sample with probabilities proportional to size (PPS), while still being able to use PRNs in order to rotate the sample. Poisson Sampling seemed appropriate, except that Poisson Sampling did not provide a fixed sample size. Around this same time Ebjorn Ohlsson (1995a, 1995b) developed Sequential Poisson Sampling (SPS), a variant of the former method with a fixed sample size, but where the probabilities of selection were merely an approximation of the desired PPS probabilities. While considering this approach for the monthly petroleum survey, Saavedra (1995) developed an improvement, which he called Odds Ratio Sequential Poisson Sampling. As it turned out, Bengt Rosen (1995) had previously not only discovered the same approach, but had demonstrated it was optimal among a class of sequential PPS sampling methods. Rosen named his approach Pareto Sampling, and it became a crucial element in the new design.

Ohlsson's method assigned a random number r between 0 and 1 and a probability p to each unit, where the sum of the probabilities added up to the desired sample size n'. Then r'=r/p became a transformed random number. Whereas Poisson sampling selected all units where r'<1, SPS selected the first n units. Pareto sampling, on the other hand defined r'=(r-rp)/(p-rp), also selecting the first n units. Pareto sampling provides a better approximation to p than SPS.

The approach that was proposed for Sample 2000 permitted exact determination of weights and applied simulations to the determination of CVs (Saavedra and Weir 1997, Weir 1997). With this approach, available resources drove the sample allocations, and weights were obtained analytically. Even more importantly, however, it was discovered that the sample 2000 design approach led to a reduction in sample size. The approach for Sample 2000 was based on Pareto sampling, as described above. In Pareto sampling, each unit is assigned a probability of selection. In this particular design, allocations were made for each cell and multiplied by each companies proportion of sales in the cell. The maximum of the measures of size across all cells became the company's measure of size.

The Sample 2000 design preserved the use of certainty companies, but modified the definition of certainty. Previous designs assigned companies that were refiners, companies that sell five percent of any one product-end use combination in a State for which estimates for that combination were needed, and

companies that had sold in more than four States to a certainty stratum. In Sample 2000, the requirement for multi-State companies to be classified as certainties was dropped. Multi-State companies that had sales of five percent or more were included by the five percent rule. In general, certainty companies reduced the total sample size because of the skewness of the distribution, but they could not be rotated out from sample to sample and, therefore, increased individual company burden.

The calculation of each company's probabilities of selection for each of the 600 potential cells was an iterative process. Initial allocations were set at the previous sample's allocation. If no estimates were required for a cell, an allocation of zero was used. Given those allocations, for each company and cell, the company's volume was 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 was then examined. If the size was too large or too small, the allocations were adjusted. This was done by preserving the certainty companies and multiplying the probability of the non-certainty companies by a constant.

The initial probabilities were used in 100 simulations. Total cell 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 original design made use of a ratio estimator, using the population in each stratum. The design for sample 2000, in comparison, did not make use of strata in sampling, but allowed strata to be defined after the fact in order to adjust the estimates. A Dalenius-Hodges procedure was applied to the non-certainty units for each target variable in each State. The strata were defined as : a) certainty, b) zero and frame non-respondents, d) low and e) high volume. The sampling methodology fixed the total sample size for the United States while sample counts within cells varied from sample to sample. Within these individual cells, the situation is similar to the variable sample size in Poisson sampling.

In any stratum, a sample expectation adjustment was made by multiplying the sample weights by E(n)/n, where E(n) is the expected sample size (equal to the sum of the probabilities of selection for all frame units in the stratum) and n is the actual sample size. This adjustment is discussed in Brewer and Hanif (1983) and replaced the population adjustment estimator from the previous design.

When the actual sample was drawn, instead of trimming the weights, the probabilities of selection were capped at .01 prior to the adjustment that set the sample size. This means that the weights ended up somewhat smaller, but extreme weights were avoided, and thus did not have to be trimmed. Since the stratification had to take place after the determination of unit probabilities, the variance estimates used during the iterative process were conducted using only the unadjusted weights.

This first cycle of the new design was the first to select companies instead of CSUs. The PRN for the home State CSU was used for each company. Using an approach similar to that used for CSUs, the companies were classified by home State and urban/rural address, and the PRNs were collocated by these categories. Finally, the rotation procedure was similar to the previous one, except that the rotating factor was proportional to the probability of selection (which was now assigned to each unit) instead of the maximum of the probabilities of strata linked by the PRN.

Use of the Chromy Algorithm

Around this time in 1997, Richard Sigman presented a talk on the Chromy (1987) allocation algorithm at the Washington Statistical Society. This algorithm is usually presented in terms of obtaining optimal allocations when multiple estimates are needed for a single stratification. It would have, in this form, reduced the allocations obtained in the early cycles of the EIA-782 when the maximum allocation among two or three end-uses was assigned to cells. The presentation of several applications of this algorithm at the 1995 Joint Statistical Meetings alerted us to its possible uses, but by then the design was moving away from linked stratified samples

Sigman's presentation, however, made it clear that the algorithm could also be used with PPS sampling by using fractional allocations and treating each sampling unit as a stratum. A fractional allocation capped by 1.0, assigned to a stratum which consists of a sampling unit is nothing more than a probability of selection. Later, after a presentation to the April 1997 Meeting of the ASA Committee on Energy Statistics (Weir, 1997), Roy Whitmore and Brenda Cox suggested that the team take another look at the Chromy allocation algorithm.

With Sigman's assistance, the Zayatz-Sigman program (1995) was used to implement the Chromy algorithm, in what may have been its first use with Pareto sampling. Thus, instead of using repeated iterations and simulations to achieve the desired allocations, the Chromy allocation algorithm was used for that purpose. Simulations were used as before to verify the adequacy of the allocations, and these provided confirmation of the results of the Chromy allocations. In fact, the sample size obtained using Chromy was similar, though slightly smaller, than the one obtained through the iterations conducted in the first cycle of Sample 2000, but its main advantage was reducing the amount of time used to complete the draw.

Sequential sampling has an additional advantage. Instead of allocating elements in the frame, it is possible to allocate respondents in-scope. This means that one can continue to sample until the desired number of respondents is obtained, and then adjust for the number of out-of-scopes and/or refusals obtained. This procedure was used for the first time in the cycle in which the Chromy allocation algorithm was first implemented.

Conclusions and Possible Improvements

There are pending issues related to the use of PRNs in this sample. One important issue is that when the frame is updated, information from several surveys which use the same frame is used to correct it. In particular, information from samples is used to separate non-respondents from out-of-scopes, and to identify problems with reported volumes in the frame. Some of these activities are beyond the control of any one survey, and so the possibility of bias is always present. However, the fact that the most critical estimates in the survey are prices, the use of adjustment factors based on the presence of non-respondents for different segments of the PRN values reduce the bias of this practice.

A second consideration is that Pareto sampling yields only an approximation of exact probabilities. Aires (1999) has developed an algorithm that calculates exact probabilities for Pareto sampling, but the fact that designated probabilities seemed to provide better results than the slightly different actual probabilities of SPS (Ohlsson, 1995a) makes this effort seem unnecessary.

In conclusion, PRNs were a part of the design of the Monthly Petroleum Product Sales Report survey over a decade prior to Ohlsson's introduction of the term. What was originally a way of reducing the required sample size of the survey, has since become a driving factor in the design. This paper chronicled the history of the use of this concept in one survey over a twenty-year period, and foresees further use of PRNs in future surveys.

References

Aires, N. (1999). Algorithms to find exact inclusion probabilities for Conditional Poisson Sampling and Pareto pps Sampling designs. *Methodology and Computing in Applied Statistics*, No. 4, pp. 463-475.

Brewer, K.R.W., Early L.J. and Joyce S.F. (1972). Selecting several samples from a single population, *Austral. J. Statist*, 14, pp. 231-239

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

Chromy, J. (1987) "Design Optimization with Multiple Objectives," *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 194-199

Ernst. L. R. (1999). The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results. International Statistical Institute, Proceedings, Invited Papers, IASS Topics, 168-182.

Holmberg, A. (2001), On the choice of strategy in unequal probability sampling, Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings, Aug. 5-9, 2001 Atlanta, American Statistical Association.

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

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

Perlman, J. and Mandel, B. (1944) The Continuous Work History Sample Under Old-Age Survivors Insurance, *Social Security Bulletin*, February

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. Proceedings, Joint Statistical Meetings, American Statistical Association, New Orleans, LA 1988.

.Saavedra, P. J. (1995). Fixed-sample-size approximations with a permanent random number. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Orlando.

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, In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Anaheim, pp. 679-682.

Saavedra, P.J., and Weir, P.. (1998). Implicit stratification and sample rotation using permanent random numbers. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Dallas, pp. 437-442.

Saavedra, P.J. and.Weir, P (1999) Application of the Chromy Allocation Algorithm with Pareto Sampling, In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Baltimore pp. 355-358

Sigman, R. (1997) "Multivariate Allocation for Stratum Sample Sizes and Poisson Sampling Probabilities". Washington Statistical Society, January 29, 1997.

Weir, P (1984) “The EIA-782B Sample Design and Estimation”. Meeting of the ASA Committee on Energy Statistics, Washington DC

Weir, P. (1997) “Data Needs-Petrolem Marketing--Sample 2000”. Meeting of the ASA Committee on Energy Statistics, Washington DC

Zayatz, L. and Sigman, R. (1995) Chromy_Gen: general-Purpose Program for Multivariate Allocation of Stratified samples Using Chromy’s Algorithm, Economic Statistical Methods Report series ESM-9502, June 1995, Bureau of the Census.

1/ The initial team consisted of Paula Weir, Mike Griffey, Bill Blackmore, Larry Thibodeau, Bob Burton, Pedro Saavedra, and Bob Clickner. Also working on the EIA-782 in its early years were David Marker, Glenn Galfond and Huseyin Goksel. Later on, Franklin Winters, Nancy Hassett, Benita O’Colmain and Richard Mantovani were at one time or another part of the team.