Application of the Chromy Allocation Algorithm with Pareto Sampling

Pedro J. Saavedra, Macro International, and Paula Weir, EIA
Pedro J. Saavedra, Macro International, 11785 Beltsville Dr., Calverton, MD 20705

PPS sampling, petroleum, simulation, permanent random numbers

The EIA-782 is a monthly price and volume petroleum product survey. The frame is a census of petroleum retailers and resalers with information on ten petroleum product/end-use combinations. The survey requires estimates at the national and regional level, as well as for all or most States for each combination. In 1997, the sample design was altered from a design which linked stratified samples to a PPS variant, using Pareto Sampling with permanent random numbers. Probabilities in 1997 were determined by repeated iterations. For the 1999 survey, the Chromy allocation algorithm was used to determine probabilities, and the Pareto sampling procedure was adjusted to account for the use of the sample in updating the frame.

1. 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 urbanization 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.

Prior to 1997, 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.

In the EIA-782 frame there are 60 geographical areas (51 States, five PADDs, three sub-PADDs and the nation). There are initially ten products of interest:

- Residential No. 2 distillate
- Nonresidential No. 2 distillate
- Wholesale No. 2 distillate
- Retail residual oil
- Wholesale residual oil
- Retail motor gasoline
- Wholesale motor gasoline
- Residential propane
- Nonresidential retail propane
- Wholesale propane

Hence there are up to 600 cells for which estimates might be required. As it turns out, estimates are not desired for each combination, but this is a policy decision which could change from cycle to cycle, or as volumetric sales shift over the years.

The new EIA-782 design, Sample 2000 as it was first implemented involved 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 was 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 was 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 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 probabilities of 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 method of selection used in that cycle was a form of order sampling (Pareto Sampling) described below.

As modified during the second cycle, the Sample 2000 design entailed two change. The first is the use of the sequential feature of Pareto sampling to continue to sample companies, dropping those found to be inactive, until the desired sample size is reached. The second is the use of the Chromy allocation algorithm to replace the iterative allocation procedure for the 600 cells.

2. Pareto Sampling

The sampling technique used for the first cycle of the new design was Pareto sampling using permanent random numbers (PRN). 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 PRN. 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 in the same way 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 0 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 for the EIA-782 and 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:

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

In other words, the product of the PRN and the probability of selection is subtracted from the numerator and the denominator used in sequential Poisson sampling. If r were itself a probability, this formula could be written as an odds ratio. Rosen demonstrated that this is an optimal order sampling method with unequal probabilities and fixed sample size. The EIA-782 uses Pareto Sampling after rotating the PRNs and collocating them by the home state of the company.

3. The Chromy Algorithm

Previously the allocations were carried out by trial and error. An initial set of allocations for each of the 600 cells (including 0 for non-publication cells) became the starting point. For each unit the proportion of the volume in the cell was multiplied by the allocation to obtain an expectation. The maximum of the expectations across cells was obtained and those units with expectations greater than one become certainties. The non-certainty probabilities were adjusted to fix the expected sample size. Then 100 to 1,000 samples (the latter figure was used at the end when greater precision is imperative) were drawn and the CVs were empirically estimated. Then cell allocations were adjusted, increasing the allocation if the CV was too large and decreasing it if it was much smaller than the target CV.

Chromy (1987) developed an algorithm to obtain an allocation for each stratum where one needs to meet multiple constraints (in our case the CVs for each publication cell). One can treat each unit as a stratum so that the resulting sample will meet a set of variance constraints. These constraints can include a maximum CV for each cell as well as a fixed sample size. Setting up the Chromy algorithm is complex, but once this is done it may yield probabilities of selection without multiple iterations. This does not replace the need for multiple iterations of some sort, since one cannot fix both sample size and CVs. Thus any CV constraints may yield a sample size that exceeds the budgetary limits of the survey, requiring the relaxation of some constraints.

Laura Zayatz and Richard Sigman (1995) of the Census Bureau developed a program in SAS which calculates probabilities for the Chromy algorithm. Most previous uses of this program have been for stratified sampling design, and indeed, the documentation refers to the program as applying to stratified samples.. However, the Chromy algorithm and the Zayatz-Sigman program permit the treatment of individual units as strata, the resulting allocations to be fractional, and thus the resulting allocations to be treated as probabilities of selection in a PPS scheme (Sigman, 1997).

The EIA-863 frame was prepared, and designated certainties were defined as reported in the Rotation Plan Report. Each company was given one record and 600 variables, one for each product and cell. Constraints were established for each variable corresponding to a publication cell. This included all gasoline and residual variables, except for those where there were no sales in the EIA-863. In addition, constraints were meaningless for cells where designated certainties accounted for 100% of the sales.

The design called for a CV of .15 for distillate publication cells and .10 for all other products. Naturally, any set of simulations would yield an average CV corresponding to the expected CV, but one could expect many to result in CVs slightly higher than the target. After some initial trials indicated that the sample size could be held comfortably below the desired number, CVs of .14 for distillate and .09 for the other products were assigned.

The parameters for the algorithm were set for PPS sampling as follows: 1) each company was treated as a a stratum, 2) its volume in each of the 600 cells became the standard deviation of a variablein the stratum, 3) the maximum allocation was set to one per stratum, and the minimum allocation was set to .01, and 4) each company was assigned an equal cost. The cost function was required to converge to .5 and ten iterations were initially allowed (fewer turned out to be required).

The result yielded approximately 2,052 companies after the inclusion of the companies with no sales reported (either 0 or missing sales). The latter companies were given (unless they were designated certainties) a probability of .01. Of these 2,052 companies, 445 were designated certainties and 72 more were certainties resulting from the allocation. In order to avoid companies with small noncertainty weight, all companies with probabilities greater than 2/3 were made certainties. This was done, adding 158 addition certainties, but only 33 more companies to the sample (since most of the 158 would be sampled in any case).

The sample was then increased slightly to 2,099 companies, of which 99 were refiners (all refiners are part of the EIA-782A, and are not counted among the 2,000

allowed by the budget). This increase was restricted to non-certainty companies with inclusion probabilities greater than .01. It is recommended that the EIA-782B sample consist of exactly 2,000 active companies.

The 675 certainties include the following:

Refiners:	99
Five per centers	264
Other designated	82
Assigned by Chromy	72
Probability > 2/3	158

Simulations were conducted for the sample, using Poisson sampling (experience shows Pareto sampling will tend to do slightly better than Poisson). Using 500 simulations, volumetric coefficients of variation were obtained for each publication cell. Not one CV exceeded the target CV for that cell (recall that the Chromy parameter was set to one percentage point lower)..

Bibliography

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

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. (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, 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.

Rosen, B. (1995) "Asymptotic Theory for Order Sampling", 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. The Use of a Variant of Poisson Sampling to Reduce Sample Size in a Multiple Product Price Survey, Proceedings of the Section on Survey Research Methods, 1997 Joint Statistical Meetings, American Statistical Association, Anaheim, California.

Sigman, R. Personal communication, 1997.

Zayatz, L. and Sigman, R. 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.

The authors wish to acknowledge the assistance of Dr. Richard Sigman, who provided technical assistance in using his program, and of Dr. Richard Mantovani and Wendy Wyatt, who served as peer reviewers for this paper.