The Use of a Variant of Poisson Sampling to Reduce Sample Size in a Multiple Product Price Survey

Pedro J. Saavedra, Macro International, Paula Weir, Energy Information Administration
Paula Weir, 1000 Independence Ave., SW, Washington, D.C. 20585

Key Words: Order sampling, PPS sampling, simulations, Permanent random numbers, EIA

Introduction

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. The basic sample design of the survey has been in effect since 1984, but has been modified in minor ways to reflect new information or changes in the market. Samples have been rotated during that time to reduce the individual company burden. The current sample is the eleventh such sample. The basic design made use of two groups, a certainty group and a noncertainty group. For each of eight 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. 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 procedure. 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 target variables, but not all, 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. 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 1000 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 this design, the desired C.V's, therefore, drove the sample allocation process, calculated separately for each basic sample.

With a reduced budget in 1996, the focus shifted to reducing survey operations' costs significantly. These costs were directly associated with the sample sizes because operational 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 2000 companies, compared to the current sample of approximately 3000 companies. This sample was, therefore, named Sample 2000. In that an operational sample 66% as large as the previous sample represented a tremendous decrease, it was expected that some variables would no longer be targeted and design CVs would be increased for other variables. Also, the requirement for state level estimates for all states, all variables, would have to be loosened to reach the reduced sample size. The various combinations of CVs for the target variables at various geographic levels that were possible to achieve a total sample of approximately 2000 were numerous. The linked selection procedure, however, did not lend itself to easy computation of individual target variable contributions to total sample size to compare the variety of scenarios. In addition, because frame data were available for the first time for one of the petroleum products by end-use type, the number of target product/end-use variables was expanded to ten.

Sample 2000 Approach and Design

The approach that was proposed for Sample 2000 permitted exact determination of weights and applied simulations to the determination of CVs. 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 itself led to a reduction in sample size. The sampling approach for sample 2000 was

based on a variation of Poisson sampling, the variant used to control sample size. In Poisson sampling, each unit is assigned a probability of selection. In this particular design, the probability was based on the volume of the unit relative to the total volume in each cell. A random number, uniformly distributed between 0 and 1 was assigned to each unit. If the random number was smaller than the probability of selection, the unit was selected. If the random number was larger, the unit was not selected for the sample.

The disadvantage of Poisson sampling is that the sample size is not fixed. However, there were three methods to limit or eliminate sample size variability. In particular, one method considered to control sample size was collocated sampling. In this method units were randomly ordered and then assigned a permanent random number, (k-x)/n where n was the number of units, k was the order of the unit, and x was a random uniformly distributed number between 0 and 1. In this new set of random numbers the smallest was between 0 and 1/n, and the second smallest was between 1/n and 2/n, etc. Collocated sample proper always uses x = .5, the midpoint of the segment. These uniformly distributed random numbers were less likely to cluster towards 0 or 1, which reduced the variability in sample size. The second method considered, developed by Ohlsson (1995), and referred to as sequential Poisson sampling, precisely controlled the sample size by order sampling. The companies were ordered by their permanent random number divided by their probability of selection. The first n of the ordered companies were selected. The third method, the Saavedra Odds Ratio Sequential Poisson Sampling (1996), or Rosen's identical procedure called Pareto sampling (1996), improved on the second method and may be shown to be optimal among a class of sampling methods. This method also used ordered sampling, but orders companies by the ratio of the permanent random number times one minus the probability to the probability times one minus the permanent random number (i.e., $r(1-p) / p(1-r)$). Again, the first n of the ordered companies were selected. This last method was incorporated in the Sample 2000 design.

A variation of collocated sampling was used, however, to resolve another design issue. In particular, it was necessary to guarantee the spread of the random numbers for noncertainty units by home State and urbanicity, as opposed to the distribution of the random numbers of the units between 0 and 1 in collocated sampling. For this purpose, two urban-rural strata per home state were used for implicit stratification. Random numbers within each home state-urban status combination were spread out so that each occupied a different segment of size 1/n, where n was the number of noncertainty companies in the home State-urban status combination. This was the same as collocated sampling, except x was random, rather than set at .5, the midpoint.

The Sample 2000 design preserved the use of certainty companies but modified the definition of certainty. Previous designs designated companies that were refiners, companies that sold five percent of any one target variable in a publication state, and companies that had sold in more than four states as certainties. In Sample 2000, the requirement for multistate companies to be classified as certainty was dropped. Multistaters were designated certainties only if they qualified by one of the other reasons. Even though, in general, certainty companies reduced the total sample size because of the skewness of the distribution, they could not be rotated out from sample to sample, which increased individual company burden. Modification of the certainty definition did release a number of companies from continued reporting without increasing the sample size and resulting total company burden. The possible disadvantage of this modification was that if the volumes from the frame were in error, the effect would be greater for companies classified as noncertainty rather than certainty. A second type of certainty, defacto certainty, also occurred in the previous design. This certainty was a company who was assigned a probability of one because the number required for sampling equaled the number in the frame for that cell. The equivalent certainty in sample 2000 was a company whose proportion of the volume in a cell multiplied by the allocation of the cell was greater than or equal to one.

The calculation of each company's probability of selection for each of the 600 potential targeted cells was derived through an iterative process. Initial allocations were set at the previous sample's allocation. If a cell was designated not 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 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. The initial total sample size was then examined. Allocations were then increased where CVs were too high, and decreased if CVs were unnecessarily low relative to the targeted size.

This allocation algorithm was modified by three rules. The first was that a downwards adjustment could not be

more than 10%. The second rule was that every adjustment added 1 to the allocation for all cells. The third rule required a minimum cell size of 10 for state allocations and 20 for higher level geographic aggregate allocations. The final probabilities were also calculated using the following rules: 1) for each cell, the proportion of volume the company sells was multiplied by the allocation, 2) the maximum of these cell probabilities for each company was assigned to each company, 3) probabilities under .01 were set to .01, 4) probabilities under 1.0 (noncertainties) were then multiplied by (2200-c)/(q-c) where c was the number of companies with values of 1.0 and q was the sum of the values of the remaining companies. Using this revised set of allocations, the process was repeated until the CVs of the estimates of frame volumes were under 15% for distillate categories and 10% for the other products. Final allocations were achieved by multiplying the previous allocations by the square of the ratio of the previous CV to the targeted CV less 1. The resulting probabilities from #4, the probabilities of selection, were then guaranteed to sum to 2200. A total sample size of 2200 was expected to result in 2000 or less active companies once implemented because of the deaths due to a dated frame.

In order to meet the goal of an operational sample size of 2000, the required 10% CVs for each state for the three propane end-use categories were relaxed. The previous sample was selected based on only whether or not the company sold propane and no frame volume data were available. In that design, the distillate sampling errors were used as a proxy for the propane errors due to the lack of propane frame data. Given that the newer frame for the sample 2000 selection contained detailed propane data, and given the resulting sampling errors realized in the previous sample, it was determined that 3200 companies would be required to meet the target CVs for all states. The decision was therefore made to target only 25 of the states for propane, as well as targeting the higher geographic aggregates. The 25 states were determined by their propane volumes for the targeted end-use categories using the frame data and industry publications. These were considered to be the most important propane states and the states that depended the most on propane. The high CVs in the other states were mostly due to propane being a minor product which yielded a small denominator in the CV calculation. One of the advantages of the new design was that it enabled the examination of options such as the use of only targeted states to produce a more useful and efficient sample overall.

Estimation Methodology

The current design made use of a ratio estimator, the sum of the sample and volume weighted prices divided by the sum of the sample weighted volumes, where sample weights were constant within a stratum that was defined on volume. The design for sample 2000, in comparison, did not explicitly stratify companies. The urbanicity strata were implicit and used to balance units within the state between urban and rural. This design, however, allows strata to be defined after sampling in order to adjust the estimates by stratum. The sampling methodology fixed the total sample size for the United States as targeted, while sample counts within the individual cells varied from sample to sample. Within these individual cells, the situation was similar to the variable sample size in Poisson sampling.

Two kinds of adjustments are appropriate for Poisson sampling and related strategies in estimating totals. These adjustments after sample selection are also effective when estimating prices. These adjustments are referred to as sample expectation adjustments and population expectation adjustments. Each of these adjustments was tested by applying a Dalenius-Hodges procedure to the noncetainty companies for each target variable in each State. Strata were defined as: a) certainty, b) zero, c) low and d) high volume.

In any stratum, the sample expectation adjustment was made by multiplying the sample weights by n_c/n, where n_c 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). Similarly, the population expectation adjustment in any stratum was made by multiplying the weights by N/N_c where N is the population of the stratum and N_c is the sum of the sampled units' weights or the population estimated from the sample. The population expectation adjustment was used in the current sample's estimation formula because of the variable sample size resulting from linked selection.

The two adjusted estimators were compared empirically for both price and volume with an approximation to the Horwitz-Thompson estimator which does not adjust the inverse of the probability of selection. (The estimator is called an approximation because the sampling approach yields a very close approximation to the designated probabilities, but the estimator made use of designated, rather than actual, probabilities of selection). While each of the adjusted estimators did better in different products, the sample expectation adjustment performed slightly better than the population expectation adjustment overall in terms of root mean square error.

In addition, to the approximate Horwitz-Thompson estimator, the weights adjusted by the sample expectation adjustment factor were tried with Poisson and Pareto sampling. The Pareto sampling yielded slightly better results in addition to offering greater control.

Implementation

Simulations on the new sample showed that median CVs across cells were one cent or under for all product/end-use categories except one. That design category is actually published at a less aggregated end-use level. It is expected that at the published level prices are more homogenous which will result in lower CVs. A few other individual, non-systematic cells exceeded the target CVs by a large amount but were found to be the result of sample versus frame data differences. These were most likely respondent reporting errors which were random and couldn't be predicted in the future sample, and were therefore ignored.

The sample 2000 was overlapped with the previous sample. For the January and February reference months, respondents for both samples were required to report. During this time new respondents were compared to old respondents and discontinuities in the data series examined and minimized. Data for the March reference month was published using only the sample 2000 respondents.

Sample rotations are scheduled annually to replace approximately 50% of the noncertainty companies. Future sample rotations, were also considered in this design. Given the new sample that was drawn, if the probabilities have not changed, the sample rotation would be accomplished by one of the following: 1) rotating the permanent random numbers (PRN) by a fixed amount, 2) rotating the PRN by an amount proportional to the probability of selection, 3) #2 and then re-parameterizing within implicit strata. The second and third methods are intended to give larger companies a better chance of rotating out of the sample. In addition, the sample 2000 design also lends itself to the ability to gradually rotate the sample at a smaller proportion more frequently. This not only spreads out potential discontinuity over time, but also has appeal as the frame frequency becomes less often in the decreasin budget environment.

Future sample rotations will consider use of Chromy's allocation algorithm (1987) which has been realized as applicable to Poisson sampling as well as to stratified sampling. The use of Chromy's algorithm is an alternative to the iterative process of sample allocation that was used in this effort.

Bibliography

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

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 Permanenet Random Numbers", Survey Methods for Business, Farms and Institutions, edited by Brenda Cox, NewYork: 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.