TWO MULTI-PHASE SURVEYS THAT COMBINE OVERLAPPING SAMPLE CYCLES AT PHASE 1

Paula Weir, Energy Information Administration and Pedro Saavedra, Macro International Inc.
Paula Weir, 1000 Independence Ave.,S.W., EI-42, Washington, D.C. 20585

KEY WORDS: Two-phase sampling; PPS sampling; Petroleum surveys; Telephone surveys

ABSTRACT

The EIA-888 is a survey of diesel fuel outlet prices that produces estimates of national and regional level prices. The EIA-878 is a survey of motor gasoline outlet prices that produces estimates of national and regional level prices, as well as separate estimates for four formulations and three grades of gasoline. Both of these weekly surveys have used a monthly survey as phase I of a multi-phase sample, subsampling the sample units of the monthly survey who report the specific outlet sales category. Recently phase I of both of the weekly surveys has used a combination of two overlapping sample cycles of the monthly survey as phase 1, adjusting the Probability Proportional to Size (PPS) size measures to account for sample units present in both sample cycles.

BACKGROUND

. EIA conducts two weekly Computer Assisted Telephone Interview surveys that collect prices at the outlet level. The first is the EIA-888 which collects prices of diesel fuel from truck stops and service stations across the country each Monday morning. The second is the EIA-878 which collects prices of regular, midgrade, and premium motor gasoline by formulation from service stations across the country each Monday morning. Average prices of gasoline and diesel fuel through outlets at the five Petroleum Allocation for Defense District (PADD) levels, regions of the country, sub-PADD levels, and the state of California are released by the end of the day through Listserv, the Web, Fax, and telephone hotline.

The diesel fuel prices that are released are used by the trucking industry to make rate adjustments in hauling contracts. Gasoline prices are frequently quoted by the media, particularly during times of rising or falling prices, because of the general interest to the public. The gasoline prices have been used in analyses of the cost of the Environmental Protection Agency regulations requiring oxygenated and reformulated gasoline in specified non-attainment areas. The prices have also been used by the state of California in helping to understand the high level of prices associated with their distinct market. Most importantly, they have provided national and state level legislators valuable independent, accurate and timely information during times of volatile markets and prevented the creation of unnecessary legislation in a free market system.

SAMPLE DESIGNS

The sample designs for these two surveys were based on the need to provide efficient samples with simple estimation to promote the fast turnaround time on gathering the data and releasing estimates. Design targets were originally set at 1 cent when the surveys began but, as more detailed information was required by customers, such as sub-PADD, grade, and formulation of gasoline, these targets were allowed to vary for lower level aggregates to provide sample sizes conducive to quick collection with minimal or no increase in survey costs. These targets are re-evaluated when new samples are drawn, historical standard errors examined, and cost-benefit analysis reviewed. The revised targets are shown in Table 1.

Table 1. TARGET STANDARD ERRORS (in cents)

Geographic Area	Diesel Fuel Target Error	Motor Gasoline Target Error All formulations (Each formulation)
U.S.	1.0	1 (1.0)
PADD I	1.0	1 (1.5)
PADD IA	1.5	1 (1.5)
PADD IB	1.5	1 (1.5)
PADD IC	1.5	1 (1.5)
PADD II	1.0	1 (1.5)
PADD III	1.0	1 (1.5)
PADD IV	1.5	1 (1.5)
PADD V	1.5	1.5(2.0[1]/1.5)
CA	1.5	1.5 (1.5)

[1] 2.0 cents standard error was targeted for conventional but 1.5 for the other individual formulations.

In addition to the timeliness requirement, the designs were driven by the lack of a frame listing of diesel fuel outlets or service station outlets. Instead, the designs made use of a monthly survey of a census of refiners, and a sample of resellers and retailers of petroleum products. This company level survey collects prices and volumes by state and enduse, and in particular, for the enduse category sales through retail outlets. The sample for this survey is rotated roughly every 12 to 18 months. Data from this survey formed the bases of the first stage of sampling for the two weekly surveys. Company-state units (CSUs) in the monthly survey with price and volume data for gasoline or diesel fuel in the sales through retail outlets categories made up the frame for the first phase sample for the weekly surveys.

DIESEL SAMPLE

To determine the allocations, average standard errors across reporting periods for the previous year of weekly diesel fuel survey prices were calculated for each of the cells. An average sample size was then determined for each cell by the formula:

$$n' = (e/t)^2 n,$$

where t was the targeted standard error, n was the previous sample size for the cell, and e the average of the previous sample's standard errors, and n' was the new sample allocation.

In addition, a second allocation based on proportional representation (proportion of diesel fuel volume sold) within the next larger cell (i.e. more aggregated level cell that the original cell would contribute to) was also obtained. For example, the PADD IB cell contributes to the PADD I and the U.S. cells. The maximum of the these two allocations for each cell was then designated as the cell allocation.

For the diesel fuel survey, data from cycle 11 of the monthly survey for November 1995 to October 1996 provided 1,536 CSUs from 964 companies. However, because in this most recent sample selection estimates were being targeted at the sub-PADD and California level for the first time, concern was raised that the increases in sample size would result in more cases of multiple outlets per CSU. In addition to increased design effects, the possibility existed of cases where the number of outlets sampled for a CSU would exceed the number of outlets that the CSU had in the particular state. As a result of these concerns, consideration was given to using data for January to June 1994 from the previous monthly survey cycle, cycle 10, thereby providing more CSUs from which to sample. Using two survey cycles of data, two separate, independent samples could be selected, one from each cycle, and outlets could be sampled from the CSUs so as not to overlap. The estimates from the two samples could then be averaged.

However, it is also apparent that there is no need to conceptualize the design as consisting of two samples. For example, consider a given CSU which is allotted a portion of the allocations in a sample. The CSU, x, has an expectation of $e_1(x)$ outlets. If the method of selection is a Goodman-Kish approach, where more than one outlet may be selected from the CSU, then the number of outlets selected will differ from $e_1(x)$ by less than one (e.g. if $e_1(x)$ =3.2 then either 3 or 4 outlets will be selected from x). Suppose the expectation for the same CSU from the second sample was $e_2(x)$. Then one could assign an expectation of $e_1(x) + e_2(x)$ to the CSU and combine the two samples into one draw.

With the combined sample cycle approach, one sample selection, the CSUs' measures of size could be normalized to sum to the cell allocations. Therefore, a proportion of the allocation could be assigned to each cycle, and each proportioned allocation could be multiplied by the proportion of weighted volume each CSU represented in the cell. Size measures could be added across cycles and only one sample selected. The results from one sample would be the same as the averaged estimates from two separate samples. The simpler one-sample method was implemeted. The volumes of companies that appeared in only one cycle of the monthly survey were multiplied by a ratio reflecting the ratio of companies present in both sample cycles. The use of the cycle 10 sample provided 1,693 CSUs from 1,089 companies. The final combined frame counts, the sample for Phase 1 for both gasoline and diesel, are provided in table 2.

The second phase had two stages. The first stage of the second phase of the sample design for the diesel fuel weekly survey used as a measure of size for PPS sampling the CSU's annual state sales volumes from the monthly survey divided by the unit's probability of selection in the monthly survey. These size measures were normalized by assigning ½ of the allocation necessary to achieve the target errors in the cell to each cycle and multiplying this half of the allocation by the proportion of the total weighted volume in the cell for the cycle represented by the CSU. The allocation procedure described above yielded a targeted sample size of 350 for the diesel fuel survey. Normalized size measures for each CSU were determined for each cycle separately, and then the two size measures were added to form one frame.

Each CSU in the frame, therefore, had a size, and the sizes of the CSUs within each cell added up to the allocation of each cell, which are shown in Table 3.

To select the units for the second phase of the sample, the frame CSUs were sorted by state and randomly ordered within each state. The normalized size measures were then used to define sampling intervals of 1.0. Using the random order, cumulative size measures were determined where a CSU's cumulative size was the sum of the sizes of all CSUs

Table 2. FRAME COUNTS FOR THE DIESEL FUEL AND GASOLINE SAMPLES (PHASE 1)

	COMPANIES		CSUs	
	Diesel Fuel	Gasoline	Diesel Fuel	Gasoline
EARLY CYCLE ONLY	567	843	671	1070
LATER CYCLE ONLY	442	605	514	743
BOTH CYCLES	522	316	1022	964
TOTAL FRAME	**1531**	**1764**	**2207**	**2777**

Table 3. DIESEL FUEL FRAME AND SAMPLE ALLOCATIONS BY CELL

PADD	1A	1B	1C	2	3	4	CA	Other	TOTAL
Phase 1: CSU Frame	150	214	325	868	296	167	44	143	2207
Phase 2: CSU Sample	22	27	53	65	35	39	18	23	282
Phase 2: Outlets	28	34	64	72	36	56	29	31	**350**

preceding it and including it. A random number between 0 and 1was chosen as a seed, and assigned to the first CSU in PADD 1A. The first CSU whose cumulative size exceeded the seed was sampled and 1.0 was added to the seed. If the CSU's cumulative size measure still exceeded the seed plus 1, the CSU was sampled again and 1 was again added. The sampling continued in this manner selecting the next CSU whose size measure exceeded the count plus seed, until the desired outlet sample size was obtained. The second stage of the second phase was to contact the sampled companies and ask them to provide outlet telephone numbers and addresses for the number of outlets in each state that the CSU was sampled. If the CSU was sampled more times than the company had outlets in that state, an outlet was counted more than once.

Since allocations were derived at the cell level, cell averages were just simple averages of the CSU prices (the weights from the first and second phases cancel). The U.S. average was a weighted average of the cell/PADD averages where the weights were derived by taking the inverse of the probability proportional to the PADD weighted volumes.

GASOLINE SAMPLE

Similarly, the gasoline sample, selected almost a year after the diesel fuel sample, made use of two frames based on cycle 11 and cycle 12 of the monthly survey as the Phase 1. In this survey, standard errors were targeted for PADDs, sub-PADDs, and California, as well as formulation. The sample sizes within the PADD/formulation (conventional, oxygenated, reformulated, and oxygenated program reformulated gasoline (OPRG)) cells were allocated using the maximum of the grades' (regular, midgrade and premium) median standard error across reporting periods for the previous 6 months of the weekly gasoline survey prices. The weekly standard errors were obtained using a bootstrap procedure. A single bootstrap covered all reference weeks, but separate variance estimates were derived for each week. Similar to the diesel fuel survey allocations, cell allocations took into account the allocation necessary for that cell itself, as well as the contribution that cell makes to a more aggregated cell by considering its proportion of total volume in the larger cell and multiplying that proportion by the allocation of the larger cell. For example, the PADD IV oxygenated gasoline considered the allocation required in that cell, as well as the proportional allocation needed in PADD IV total for all formulations of gasoline and the U.S. total oxygenated gasoline allocation. The maximum of the allocations was assigned to the smaller cell to satisfy all requirements. However, because the first stage of the second phase sample yields company-state units, and companies do not always have available a list of outlets designated by attainment status (i.e. formulation), the number of outlets originally sampled from each CSU often had to be larger than the number actually desired in order to satisfy the individual formulation allocations. Once the attainment status of the outlets was determined during initiation of the sample, the desired number of outlets could be subsampled to obtain the targeted sample size .

To produce CSU expectations on the number of outlets required in oversampling to achieve the desired number of outlets for each formulation, ratios of the formulations were derived using the monthly survey where possible. Where ratios could not be calculated, such as ones involving OPRG, which is not collected separately in the monthly survey, population ratios for the specific attainment status at the state level were used instead. For each CSU, the CSU's total gasoline volume across the three grades was multiplied by the expected proportion of gasoline for each of the four formulations to yield an expected volume by formulation. The CSU's monthly weight was applied to these formulation volumes and divided by the total weighted volumes for the PADD for that formulation and then multiplied by the cell's desired allocation to yield the expected number of outlets to be sampled for each CSU. This was done separately for companies in each of the two monthly respondent cycles and the results of the two cycles added together. These expectations were then divided by their proportion of the CSU volumes. The maximum of these four results, one per formulation, was the global expectation or the size measure that was used for each CSU. Sampling then proceeded as in the diesel fuel sample, using Probability Proportional to Size (PPS) and a sampling interval of 1. Because of the use of oversampling, second stage sampling was necessary. For each outlet selected, the outlet's adjusted expectation was divided by the maximum adjusted expectation. If this quotient was larger than a selected random number between zero and one, the outlet was retained. If the quotient was smaller, the outlet was dropped.

The second phase sampling produced 507 CSUs from 304 companies. Stage 1 resulted in 1,174 outlets, and stage 2 yielded an expectation of 820 outlets as shown in Table 4. Final stage 2 expected and actual allocations by formulation are shown in Table 5.

Table 4. GASOLINE FRAME AND SAMPLE ALLOCATIONS (1ST STAGE AND 2ND STAGE EXPECTATIONS) BY CELL

PADD	IA	IB	IC	II	III	IV	CA	Other V	TOTAL
Phase 1: CSU Frame	176	238	373	1165	337	253	48	187	2777
Phase 2: CSU Sample	52	58	62	123	55	61	20	75	507
Phase 2: Outlets 1st Stage	164	171	71	211	212	124	38	183	1174
Phase 2: Expected Outlets 2nd stage	64	103	65	175	85	124	38	166	**820**

Table 5. GASOLINE SAMPLE SECOND PHASE SECOND STAGE OUTLET COUNTS: EXPECTED AND ACTUAL BY FORMULATION

PADD	Total 2nd Stage		Conventional		Oxygenated		Reformulated		OPRG	
	Exp.	Act.	Exp.	Act.	Exp.	Act.	Exp.	Act.	Exp.	Act.
US	820	783	406	388	163	157	177	173	74	65
IA	64	63	15	15	0	0	34	33	15	15
IB	103	92	31	25	0	0	14	21	58	46
IC	65	60	52	54	0	0	13	6	0	0
II	175	178	86	82	42	55	47	44	0	0
III	85	78	34	33	21	14	30	31	0	0
IV	124	122	75	74	49	48	0	0	0	0
CA	38	38	0	0	0	0	38	38	0	0
Other V	166	152	113	105	51	43	1	0	1	4