A SIMPLE APPROXIMATE VARIANCE ESTIMATE FOR A COMPLEX UNEQUAL PROBABILITY SAMPLE SURVEY[1/]

Glenn J. Galfond, Mary E. Garvin, James C. Thompson
Price Waterhouse, Washington, D.C.

Abstract

The Energy Information Administration (EIA) annually collects sales volume data for petroleum products on the EIA-821 survey. The EIA-821 sample is a stratified random sample of wholesale and retail petroleum product dealers with unequal sample selection probabilities within strata. Volume is estimated using an inverse-probability weighted ratio estimate. Variance estimation is made difficult by the complexity of the sample design and by the fact that joint sample selection probabilities are unknown.

A simple approximate variance estimate that does not rely on joint sample selection probabilities is proposed and evaluated using simulation. The estimate is shown to be nearly unbiased and, compared to the variance estimate used in the past by the EIA, to provide a substantial reduction in bias and root-mean-square-error for variance estimation.

Background on the EIA-821 Survey

The EIA-821 survey annually collects end-use sales data from a sample of kerosene, distillate fuel oil and residual fuel oil dealers. Establishments in the sample report product volume sales for each state in which they operate by product and end use. Volume estimates calculated from the EIA-821 survey are published by state, product, and end-use in the Petroleum Marketing Monthly [1].

The sample design for the EIA-821 survey includes stratification by state, product, end use, and sales volume within state/ product/end use. A single establishment may be included in many strata, for example as a large seller of retail residual fuel oil in one state and a small seller of wholesale distillate fuel oil in another state. In order to reduce respondent burden and survey processing costs, the EIA-821 sample is designed to achieve sufficient accuracy in all target state/product/end use volume estimates while reducing the total number of establishments in the sample. Background information on the sample design and estimation procedures is presented below.

Stratification

The EIA-821 sample design is based on 51x5=255 simultaneous stratifications of the same population of establishments. The 255 separate stratifications correspond to combinations of 51 states (the 51st state being the District of Columbia) and 5 major product/end use groupings: (1) residential retail distillate, (2) nonresidential retail distillate, (3) wholesale distillate, (4) retail residual, and (5) wholesale residual.

For each of the 255 state/product/end use combinations, the population of relevant establishments (i.e., those establishments known to sell the product for the end use in the state) is stratified by sales volume. An establishment in one state/ product/end use/sales volume stratum may also be in many other strata relating to other states, products, end uses, or sales volume categories. However, within any one of the 255 state/product/end use target cells, each relevant establishment belongs to exactly one sales volume stratum.

Sample Selection

Minimum sample sizes for each stratum are calculated using frame data to achieve desired levels of accuracy in the volume estimate for each state/product/end use target cell. Rather than select samples independently for each stratum, a "linked sample selection" procedure is used to select a sample of establishments that simultaneously satisfies the minimum sample size requirement for all strata while reducing the total number of establishments in the sample.

The linked sample selection procedure is based on an iterative process of selecting establishments from a randomly sorted list of all establishments. The first establishment on the randomly sorted list is selected into the sample. The sample count for each stratum in which the establishment is a member is set to one. Next, the second establishment on the randomly sorted list is selected into the sample and the sample count for each stratum in which the establishment is a member is incremented by one. The process continues from top to bottom of the randomly sorted list. When an establishment is chosen from the list and the sample size requirement for all strata to which the establishment belongs has already been satisfied, the

establishment is excluded from the sample. Eventually, the sample size requirements for all strata are satisfied and the sample is complete.

Basic and Volunteer Sample Units

The sample selected for an individual stratum includes two distinctively different types of establishments, referred to as basic and volunteer sample units. Basic sample units for a particular stratum are those establishments selected into the sample to satisfy the minimum sample size requirement for that stratum. Volunteer sample units for a stratum are those establishments selected into the sample after the stratum's minimum sample size requirement was satisfied. Volunteer units are included in the sample for the stratum because they contributed to satisfying the minimum sample size requirement for some other stratum (i.e., they are a basic sample unit for some other stratum). An individual sampled establishment can simultaneously be a basic sample unit in one or more strata and a volunteer sample unit in zero or more strata.

Basic and volunteer sample units differ in terms of sample selection probabilities and volume distribution. Basic sample units represent an equal probability sample within a stratum, each with sample selection probability equal to the minimum sample size requirement for the stratum divided by the frame count for the stratum. In contrast, volunteer sample units represent an unequal probability sample with sample selection probabilities depending in a complex manner on attributes of other strata for which they are a member. It is believed that volunteer sample units may also tend to have larger volumes than basic sample units. This is because volunteer sample units tend to be establishments that do business in many states and sell many products and end uses, and consequently may tend to be larger companies.

Volume Estimation

The volume estimate for a state/product/end use target cell is based on the relevant design-level stratification for that cell. For example, the published estimate for total residential retail sales of distillate fuel oil in New York is based on one of the 255 design-level stratifications while the estimate of total nonresidential retail sales of distillate fuel oil in New Jersey is based on another. Separate estimates for total volume within each state/product/end use/sales volume stratum are computed and then aggregated over sales volume strata to compute total state/product/end use volume estimates.

In general, the sample in each stratum can include both basic and volunteer units, and is thus an unequal probability sample. A large scale simulation of the sample design is used to estimate the probability that each establishment is selected in the overall EIA-821 sample, i.e., selected in at least one of the 255 design-level stratifications. This sample selection probability is computed at the establishment level, ignoring the distinction between basic and volunteer units. (An alternative approach in which sample selection probabilities are computed separately for basic and volunteer units is not used because it greatly increases processing complexity. While the current approach requires computing only one sample selection probability for each establishment, the alternative approach would require computing sample selection probabilities for each establishment in as many as each of the 255 stratifications.)

Total volume for each stratum is estimated using an inverse probability weighted ratio estimate:

$$(1) \quad \hat{V} = N * \mathrm{Sum}_i(W_i V_i) \; / \; \mathrm{Sum}_i(W_i)$$

where the summation is over all basic and volunteer sample units in the stratum and:

N = stratum population count

V_i = volume reported by sample unit

W_i = sampling weight for sample unit i (equal to the inverse of the sample selection probability)

Stratum-level volume estimates are then aggregated across sales volume strata within state/product/end use target cells. (To simplify notation, we have not included a subscript to denote stratum. However, all equations in this paper implicitly refer to observations within a single stratum.)

Prior Approach to Variance Estimation

In the past, the EIA has estimated the variance of the volume estimate by the variance that would have resulted had volume been estimated using only the basic sample units. The primary motivation for such an approach is simplicity. Since, as mentioned above, the basic sample is an equal probability sample within strata, standard stratified random sampling theory provides simple variance estimates. Presumably such an approach will overstate the true variance of the

volume estimate because including the volunteer sample units in the volume estimate increases sample size and reduces variance. However, the EIA would prefer to be conservative and overstate variance when no unbiased variance estimate is available.

Since the basic sample units are an equal probability sample within strata, the stratum-level volume estimate that would result from using only the basic sample is:

$$\hat{V}_{bas} = N * Sum_b(V_b) / n_{bas}$$

where the summation is over only the basic sample units in the stratum and:

n_{bas} = number of basic respondents in the stratum

The variance estimate for the above volume estimate is:

(2) Basic Variance Estimate =

$$S^2(bas) * (1 - n_{bas}/N) * N^2 / n_{bas}$$

where:

$S^2(bas)$ = sample variance of the basic volumes in the stratum

Alternative Variance Estimate

A review of variances estimated using the Basic variance estimate suggested that the estimate significantly overstated variance in strata with a large number of volunteer sample units. Research was therefore undertaken to determine if an alternative variance estimate might reasonably measure the reduction in estimate variability attributable to the volunteer sample units.

For a single stratum, the volume estimate $\hat{V}$ in Equation (1) has mean square error:

$$MSE(\hat{V}) = N^2 * E[Sum_i(W_iV_i) / Sum_i(W_i) - R]^2 =$$

$$N^2 * E[(Sum_i(W_iV_i) - R*Sum_i(W_i)) / Sum_i(W_i)]^2 =$$

$$N^2 * E[Sum_i(D_i) / Sum_i(W_i)]^2$$

where

R = Average population volume

$D_i = W_iV_i - R*W_i$

Retaining the first term of the Taylor series expansion around the denominator $E[Sum_i(W_i)]$ in a manner analogous to that used in equation 5.5, Theorem 5.3 of [2], the Taylor series approximation to the mean square error of the ratio estimate is:

Taylor Series Approximation =

$$N^2 * E[Sum_i(D_i)]^2 / (E[Sum_i(W_i)])^2$$

Since W_i is equal to the inverse of the probability that population unit i is in the sample, $E[Sum_i(W_i)]$ equals N and the above equation reduces to

(3) Taylor Series Approximation =

$$E[Sum_i(D_i)]^2$$

Equation (3) above can be expressed in terms of population volumes and marginal and joint sample selection probabilities using the traditional Horvitz-Thompson variance estimate for unequal probability samples [3]. Unfortunately, the joint sample selection probabilities for the EIA-821 survey are unknown. A very large simulation is currently required to estimate marginal sample selection probabilities, and an even larger simulation to estimate joint sample selection probabilities is impractical.

While an exact expression for the variance of an unequal probability sample from a finite population depends on joint sample selection probabilities, it may be possible in some cases to approximate that variance without using the joint sample selection probabilities. One approximation that was tested and proved viable was to approximate the effect of unequal probability sampling from a finite population using the finite population correction factor for equal probability samples. This approach estimates the mean square error of the volume estimate as:

Both Ratio Variance Estimate =

$$S^2(D_i') * n_{bv} * (1 - n_{bv}/N)$$

where:

$S^2(D_i')$ = sample variance of D_i' (including both basic and volunteer units, with denoninator $n_{bv}-1$)

$D_i' = W_iV_i - R'W_i$

$R' = Sum_i(W_iV_i) / Sum_i(W_i)$

n_{bv} = number of basic and volunteer respondents.

The above variance estimate is referred to as the Both Ratio variance estimate because it uses both the basic and volunteer sample units and it relies on the Taylor series approximation for the variance of ratio estimates. The Both Ratio variance estimate is proposed to replace the Basic variance estimate used in the past by the EIA.

Analysis Methodology and Results

Two separate simulation studies were used to evaluate the prior (Basic) and proposed (Both Ratio) variance estimates. The first study was a bootstrap simulation based on volume reported to the 1984 EIA-821 survey. These responses are, of course, based on a sample of establishments on the EIA-821 frame. The second study was a simulation based on all responses to the initial frame survey upon which the EIA-821 sample design was based. This frame survey, the 1981 EIA-764, was a census of all known petroleum product dealers.

Each of the two simulation studies has certain advantages over the other. The bootstrap simulation relies on more recent data than the frame simulation, and may therefore provide a better evaluation of the magnitude of differences between the two variance estimates. In particular, the bootstrap simulation better represents the variance associated with "zero" and "nonresponse" strata which are composed of establishments that either reported zero volume to the frame survey or failed to respond to the frame survey. However, the bootstrap simulation can only approximate the true distribution of population volumes and the effect of the linked sample selection procedure. In contrast, the frame simulation captures the exact distribution (although somewhat out of date) of population volumes and all of the intricacies of the sample design.

The methodology and results of the two simulation studies are described separately below. Since the results of the two studies were very similar, we only summarize the results of the bootstrap simulation and provide detailed results for the frame simulation.

Bootstrap Simulation Methodology and Results

The bootstrap simulation is based on a pseudopopulation created from sample data to approximate the true population of establishments. The pseudopopulation was created by replicating each respondent observation to the 1984 EIA-821 survey based on the observation's sampling weight. For example, observations for an establishment with weight 2.0 were replicated two times and observations for an establishment with weight 2.25 were replicated either two or three times with probabilities 0.75 and 0.25 respectively.

Random samples of basic and volunteer units for each stratum were repeatedly drawn from the pseudopopulation. Since the joint sample selection probabilities induced by multistate and multiproduct establishments could not be determined, the sample for each stratum was selected independently. First, a basic sample of size equal to the minimum sample size for the stratum was selected at random with equal probability. Next, the volunteer sample from each stratum was selected based on a random "coin toss" for each member of the stratum pseudopopulation that was not included in the basic sample. The probability that an establishment was selected as a volunteer unit in a stratum was equal to the conditional probability that the establishment be in the sample given that it was not sampled as a basic unit. This conditional probability was computed based on the unconditional probability that an establishment is sampled (the inverse of the sampling weight) and the probability that the establishment is selected in the basic sample (the stratum minimum sample size divided by stratum frame count). The outcomes of the volume and variance estimates under each of 1,000 repetitions of the random sample were computed and the results across repetitions were tabulated. Results of the simulation indicated that the Basic variance estimate had a large positive bias and that the proposed Both Ratio variance estimate was nearly unbiased. The results also indicated that the Both Ratio variance estimate was substantially less variable than the Basic variance estimate.

Frame Simulation Methodology and Results

After the bootstrap simulation provided favorable results, a larger scale simulation based on survey frame data was conducted. The frame simulation involved selecting 1,000 random samples of respondents to the 1981 EIA-764 frame survey according to the sample design for the EIA-821 survey. Volume and variance estimates for each state, product, and end use were calculated for each of the 1,000 random samples based on volumes reported to the EIA-764 survey. The variance of the volume estimates across the 1,000 samples was used to represent the "true" variance of the volume estimate for our comparisons. The average and standard deviation of the Basic and Both Ratio variance estimates across the 1,000 samples was computed and used to estimate bias and root mean square error.

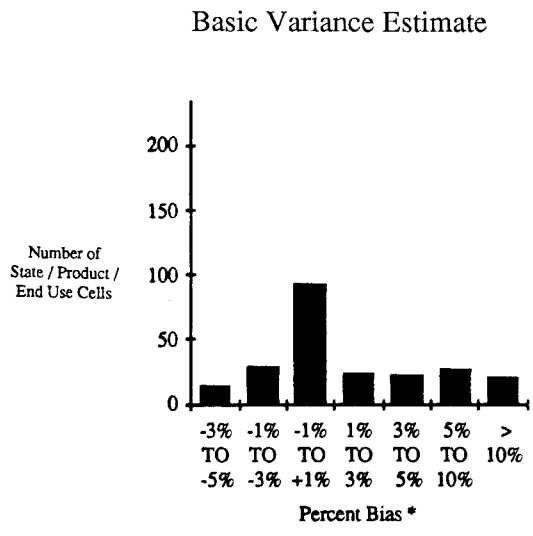
The left and right hand histograms in Exhibit I summarize the simulated bias of the Basic and Both Ratio variance estimates respectively across 232 state/product/end use target cells (23 of the target cells had 100 percent sampling rates and were excluded from the frame simulation). In Exhibit I, bias is defined as the difference between (a) the square root of the average simulated variance estimate divided by the simulated population volume and (b) the simulated coefficient of variation (CV) of the volume estimate. The biases in Exhibit I are based on CVs expressed as percentages, and so the difference between a 5% estimate and a 4% true CV appears as a 1% bias.

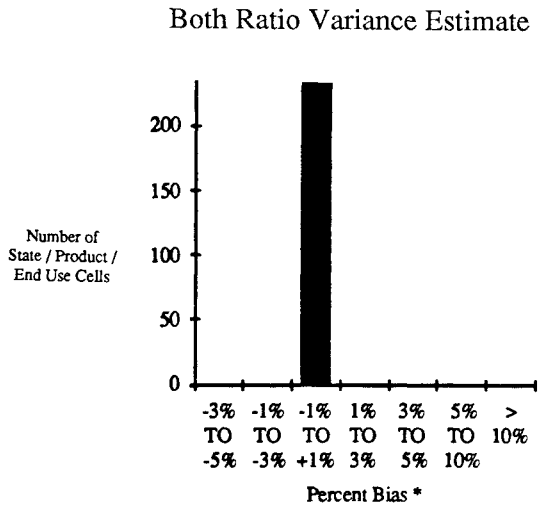
A cursory review of Exhibit I indicates that the Both Ratio variance estimate is less biased than the Basic variance estimate. While the bias of the Basic variance estimate is frequently three or more percent, the bias of the Both Ratio variance estimate was less than one percent for all of the 232 state/product/end use target cells.

Exhibit II summarizes the relative efficiency of the Basic variance estimate as compared to the Both Ratio variance estimate, where relative efficiency is defined as the ratio of their respective root mean square errors (RMSE). Exhibit II demonstrates that the Both Ratio variance estimate usually has a substantially lower RMSE than the Basic variance estimate, i.e., the ratios of RMSE's are substantially smaller than 100 percent for most state/product/end use cells.

EXHIBIT I

BIAS OF EIA-821 COEFFICIENT OF VARIATION ESTIMATES

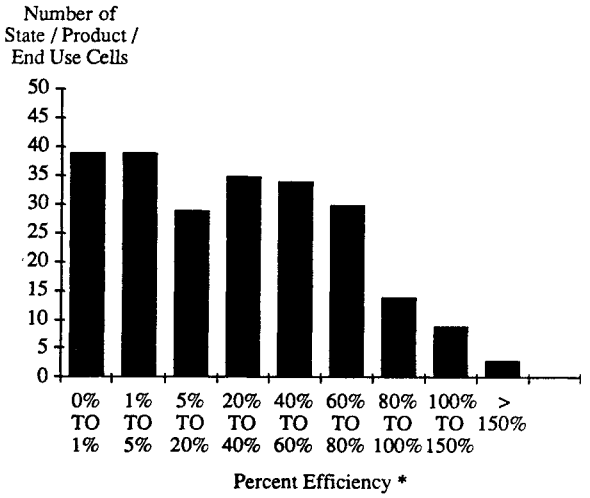




* Square root of average variance estimate estimate divided by simulated population volume, less simulated true C.V. of volume estimate.

EXHIBIT II

RELATIVE EFFICIENCY OF EIA-821 BASIC VARIANCE ESTIMATE AS COMPARED TO BOTH RATIO VARIANCE ESTIMATE



* Root Mean Square Error of Both Ratio variance estimate divided by that of Basic variance estimate.

Conclusions

The simulation studies demonstrate that the Both Ratio variance estimate provides a substantial reduction in bias and variability for EIA-821 survey as compared to the prior approach. As a result, the EIA is adopting the Both Ratio variance estimate in its next reporting of EIA-821 survey results.

The fact that the Both Ratio variance estimate is nearly unbiased suggests that, in some cases, the finite population correction factor for equal probability samples may provide an accurate approximation of the effect of a finite population on the variance of estimates computed from unequal probability samples.

References

[1] Petroleum Marketing Monthly, July, 1986, DOE/EIA-0380(86/07), U.S. Government Printing Office, Washington, D.C.

[2] Raj, Des, Sampling Theory, p. 89, McGraw-Hill, New York, 1968.

[3] Raj, Ibid., pp. 52-53.

1/ The research described in this paper was performed under contract to the U.S. Department of Energy, Energy Information Administration.