VARIANCE ESTIMATION FOR EIA-877 PROPANE PRICES

Benita J. O'Colmain, Pedro J. Saavedra, Paula Weir, Macro International Inc.
Benita O'Colmain, 11785 Beltsville Drive, Calverton, MD 20705

KEY WORDS: Survey, Systematic Sample, Error, Fuel

The Energy Information Administration (EIA) receives No. 2 fuel oil and propane price data from various State Energy Offices on a semi-monthly basis during the winter months as part of the EIA-877 State Heating Oil and Propane Program (SHOPP). The individual State Energy Offices telephone a sample of propane and heating oil companies and request current price data for each reporting period. State and regional aggregate price estimates are published in the EIA Winter Fuels Report.

The price estimates derived from the survey are subject to sampling error. The heating oil sample is drawn using a stratified random sampling technique and variances of heating oil price estimates are computed using a classic stratified random sample variance formula. The propane sample is drawn using a stratified systematic sampling technique and requires a modification to the classic variance formula. This paper describes the proposed variance estimator for propane prices and presents the results of tests using this formula in comparison with the classic variance formula.

Theory

The EIA-877 uses a stratified sampling methodology with each state forming a separate stratum. For sampling purposes propane outlets are classified as either certainty or noncertainty. Certainty outlets are those which belong to a company with outlets in more than one state or a company with more than one outlet in any one state. At least one outlet in each state where these companies sell propane is selected into the sample. The certainty outlets are further divided into two types based on the number of outlets selected for each state. First, certainty companies for which more than one outlet has been sampled for a particular state are ordered by zip code, and sampled systematically; each sampled outlet is assigned the same weight which is the inverse of the proportion of outlets selected to the total number of outlets for that state. Second, certainty companies for which exactly one outlet is selected are sampled randomly and assigned a weight which is the total number of outlets in the sampled state. The noncertainty outlets are ordered by zip code and a systematic sample (sampling every kth outlet) is drawn, again each outlet is assigned the same weight which is the inverse of the proportion of the number of non-certainty outlets drawn to the total number of outlets for that state.

Ignoring the fact that outlets are sampled systematically, each certainty company can be treated as a separate stratum, and the combined non-certainties as a stratum of their own. Then a stratum can be defined in three ways: 1) as a certainty company for which two or more outlets have been sampled, 2) as a pair of certainty companies for which one outlet was sampled for each. (These outlets are ordered by weight and paired to form a stratum, with the possibility of three companies making up the last stratum), and 3) the combination of all non-certainty outlets in the state. This would yield a stratified sample in the classic sense, but if the systematic sampling was effective, it would yield an overestimate of the variance. We will first describe the variance for the classic approach, and discuss two alternatives that take into account the systematic sampling for stratum types 1 and 3.

The uncorrected approach yields a classic stratified sample variance estimate. The variance of the weighted average price for these outlets is calculated first by obtaining the unit variance, S_k^2 for stratum k in each state:

$$S_k^2 = S_{kq}^2 + \hat{P}^2 S_{kv}^2 - 2\hat{P}S_{kqv}^2 \qquad (1)$$

where:

$$S_{kq}^2 = \frac{\sum_{i=1}^{n_k} (P_{ik}V_{ik}-\overline{P_kV_k})^2}{n_k-1}$$

$$S_{kv}^2 = \frac{\sum_{i=1}^{n_k} (V_{ik}-\overline{V_k})^2}{n_k-1}$$

$$S_{kqv}^2 = \frac{\sum_{i=1}^{n_k} (P_{ik}V_{ik}-\overline{P_kV_k})(V_{ik}-\overline{V_k})}{n_k-1}$$

n_k = number of respondents in stratum k

N_k = number of population units in stratum k

V_{ik} = reported volume for unit i in stratum k

$\overline{V_k}$ = average volume for sample units in stratum k

$P_{ik}V_{ik}$ = reported revenue for unit i in stratum k

$\overline{P_kV_k}$ = average revenue for sample units in stratum k

$\hat{P}$ = weighted average price for each State

Now, having obtained this unit variance, the variance of the State-level price estimate can be calculated as follows:

$$VAR(\hat{P}) = \frac{1}{\hat{V}^2} \sum_k N_k^2 \left(\frac{1-f_k}{n_k}\right) S_k^2 \qquad (2)$$

where f_k is n_k/N_k, the sampling fraction and $\hat{V}$ is the aggregated weighted volume for each state.

This estimate completely ignores the systematic nature of the sample. There is, in fact, no way of obtaining an unbiased estimator for a systematic sample, in that there is no way of knowing that every kth unit is not different from the preceding ones. This would be the case if one were to be sampling homes in order and every kth unit were to correspond to a corner house. But, there is no reason to suppose that ordering by zip code yields this type of result. Indeed, there is reason to believe that units with proximate zip codes are alike, thus the systematic sampling procedure would insure heterogeneity, and thus representability of the sample.

One approach to adjust for the systematic sampling procedure is to treat the noncertainty sample as a stratified sample where the stratum size is 2k and every kth unit is selected. If an odd number of noncertainty units were selected, then the last should be made part of a stratum with three units selected. While this does not correct the estimate completely, it takes into account the systematic nature of the sample. It is certainly an improvement over the classic estimator.

The recommended approach is derived from a communication from K.R.W. Brewer. Brewer derived the notion from a question regarding unequal probability sampling, but the suggestion applies to the equal probability sample as well. What is suggested here is not specifically Brewer's method, but is derived from his suggestion. The standard unit variance through algebraic manipulation can be expressed alternatively as:

$$S^2 = \sum_{i=1}^{n} \frac{(x_i-\overline{X})^2}{n-1} = \sum_{i<j}\sum_{j=2}^{n} \frac{(x_i-x_j)^2}{n(n-1)}$$

Thus the unit variance may be represented as an average square deviation of pairs of units in the sample. Now the systematic sample is effective to the extent that adjoining units are more homogeneous than units separated in the order. Thus, the unit variance can be estimated strictly from adjoining pairs of units, resulting in the formula:

$$S^2 = \sum_{i=1}^{n-1} \frac{(x_i-x_{i+1})^2}{2(n-1)} \qquad (3)$$

The rest of the formula can remain the same, and there is no need for further stratification. The same correction can also be applied to the multi-outlet certainty strata, resulting in the standard formula when only two units are sampled, but a corrected formula for three or more. Substituting the corrected unit variance as shown in equation (3) into equation (1) the following terms are derived:

$$S_{kq}^2 = \frac{\sum_{i=1}^{n_k-1} (P_{ik}V_{ik}-P_{i+1,k}V_{i+1,k})^2}{2(n_k-1)}$$

$$S_{kv}^2 = \frac{\sum_{i=1}^{n_k-1} (V_{ik}-V_{i+1,k})^2}{2(n_k-1)}$$

$$S_{kqv}^2 = \frac{\sum_{i=1}^{n_k-1} (P_{ik}V_{ik}-P_{i+1,k}V_{i+1,k})(V_{ik}-V_{i+1,k})}{2(n_k-1)}$$

Results of Testing

The two variance formulas (classic and corrected) were tested using reported propane price data for the 16 reference periods from October, 1992 through April, 1993. The variance of propane prices for noncertainties only was computed using each formula. The resulting standard deviations for each method were compared at the State level and overall using a matched pair T-test. The results of this analysis are shown in Table 1. A negative T-value indicates a lower variance for the corrected formula as compared to the classic formula. A total of 13 of the 22 States showed a lower variance with the corrected variance estimator. All differences for these States were statistically significant at the .0001 level with the exception of Kansas which was significant at the .05 level and Iowa which was not significant. The remaining nine States showed a higher variance using the corrected method. The differences were not significant for two States and the T-values for the remaining seven States were significant at the .001 level. The overall T-value was negative and significant at the .001 level.

To further examine the cause of the differences, the variance was computed using the same price data but holding the volume constant for all companies in a State. Since the price estimate is weighted by both volume and sampling weight, the variance of the volume also contributes to the overall variance. The variance computed while holding the volume constant for each company gives a better indication of whether the reported prices are behaving as expected in the systematic sampling approach. The results of this analysis are shown in Table 1. This analysis also supports the use of the corrected variance estimator. The difference between the two estimators for price variation alone seems to be more significant when the corrected estimator is lower and less significant when the corrected estimator is higher as indicated by higher T-values (lower if the T-value is negative) in most States and a lower overall T-value.

Conclusion

The results of the analysis indicate that the assumption of a lower variance using the corrected formula did hold true for some States and overall. This was more evident when holding the volume constant as was shown in the second table. The analysis also suggests that company prices in those States where the corrected formula yielded a higher variance be examined to determine whether the systematic sampling technique is appropriate for those States. The corrected variance formula was implemented to calculate the variance of propane prices for the SHOPP survey for the following survey year.

Reference

Brewer, K. R. F. and Hanif, M., *Sampling with Unequal Probabilities*, 1983, Springer-Verlog New York Inc.

Table 1. Matched Pair T-Values and Mean Differences Between the Standard Deviations for the Two Variance Estimators by State and Overall.

STATE	PRICE AND VOLUME		PRICE	
	Mean Difference[1]	T-Value	Mean Difference[1]	T-Value
Connecticut	0.358	9.70*	0.197	9.08*
Delaware	0.058	1.83	0.303	6.28*
Iowa	-0.017	-0.74	-0.085	-4.08*
Indiana	0.057	1.80	0.012	0.32
Kansas	-0.073	-2.07**	0.050	2.33**
Massachusetts	0.485	103.27*	0.107	2.70**
Maryland	0.148	3.92*	-0.027	-0.62
Maine	-0.573	-28.92*	-0.240	-11.18*
Michigan	0.343	9.94*	-0.183	-6.85*
Minnesota	-0.285	-5.88*	-0.758	-10.71*
Missouri	-0.056	-2.99**	-0.133	-4.02*
North Carolina	0.677	23.70*	0.699	34.85*
North Dakota	-0.354	-9.90*	-0.470	-10.76*
New Hampshire	-0.963	-25.57*	-0.728	-25.17*
New Jersey	-1.121	-75.71*	-0.401	-28.17*
New York	-1.140	-34.21*	-0.491	-18.82*
Ohio	-0.550	-7.03*	-0.794	-9.80*
Pennsylvania	0.700	35.26*	-0.285	-7.37*
South Dakota	-0.249	-5.01*	-0.397	-6.51*
Virginia	-1.216	-34.19*	-0.466	-14.78*
Vermont	0.337	13.63*	0.155	6.92*
Wisconsin	-0.119	-4.04*	0.049	1.88
OVERALL	-0.162	-5.26*	-0.181	-8.66*

[1] The difference is found by subtracting the standard deviation computed using the classic formula from the standard deviation computed using the corrected formula. The mean difference is the average for the 16 reference periods.

* Significant at the .001 level
** Significant at the .05 level