RATIO ESTIMATION AND APPROXIMATE OPTIMUM STRATIFICATION IN ELECTRIC POWER SURVEYS

James R. Knaub, Jr., U.S. Department of Energy

Energy Information Administration, EI-541, Washington, DC 20585

Key words:
Optimum allocation, coefficient of variation (cv), Keyfitz method, double ratio estimate

Abstract:
Some of the history of a monthly sales and revenue survey, and the current status of that survey and others, are given as background to changes now being implemented. Sampling will now take place at the State level, instead of the national level; more complete use will be made of auxiliary information; work has been done to control cv's for several variables simultaneously; and a test originally designed to consider degrees of homogeneity between several similar populations with limited sampling, has been used to investigate the effectiveness of this design at an aggregated (Census division) level.

1. INTRODUCTION:

A relatively small number of utilities sell most of the retail electricity in the United States. The Energy Information Administration (EIA) Form EIA-826, "Monthly Electric Utility Sales and Revenue Report With State Distributions," collects data from these large utilities and a stratified random sample of the smaller utilities. The universe is from another survey, begun in 1984, which is a census, performed on an annual basis.

Until 1986, the predecessors to Form EIA-826 gathered data from only a select group of large privately owned utilities. Stratified random sampling has been used with good results since that time. A "certainty" stratum has been used to let the largest utilities represent only themselves, and this design has performed well.

Currently, the strata from which random sampling takes place were established at the national level. Thus, sales and price estimates could not be published at the State level without making the assumption that the unrepresented small utilities in a State behave like (in terms of percent changes in revenue and sales volumes overtime) the larger utilities in that State. Further, State-level sales and price cv's could not be published without making horrendous assumptions. State-level sales numbers have been published, and when accumulated for a year and compared with the census when it became available, these estimates performed well. This is because approximately 80 percent or more of the sales were contained in the certainty stratum for most States and class of service "sectors." Although revenues were not considered to such a large degree, correlation between sales and revenue, which generally makes price less variable than sales or revenue, would make monthly price estimates of value also.

In the fall of 1988, the author investigated the possibility of sampling at the State level, without increasing respondent burden to an unacceptable level, by use of a nominal reduction in coverage in the certainty stratum, and by not requiring a utility selected by reason of business it does in one State to report their business in another State. That is, the unit of response is to be a utility-State part, not total utility. However, most utilities only do business in one State, so this change does not have an impact on them.

Initially, it appeared that sampling would not be possible at the State level, but could be accomplished at the "Census division" level. Example calculations of upper bounds for Census division level cv's (i.e., without auxiliary data) were promising. (There are ten Census divisions within the United States, as used by the EIA.) If it were determined that the small utilities behave more like the other small utilities in the same Census division than the larger utilities in the same State, estimates could be adjusted accordingly. Current State-level sales estimates are calculated in a manner related to the combined ratio method, but with truncated terms in the numerator and denominator. The EIA has reported, as "expansion factors," the ratio of all sales for each State and class of service sector, for all utilities in that part of the frame, to the sales for only those utilities in that same part of the frame that were also on the list of EIA-826 survey respondents. These factors have been used to upwardly adjust the otherwise unweighted total of EIA-826 responses in each State and sector. However, no such direct use was made of auxiliary data in calculating price estimates. The estimators proposed by the author make more standard use of auxiliary data for both sales and price estimation.

Early in 1989, the EIA decided to drop some of the reporting requirements on the Form EIA-826 which are not pertinent to the main objectives of this survey. This reduced calculated burden per respondent enough to allow sampling at the State level with a total calculated respondent burden approximately unchanged from the previous year.

With the existence of an annual census to supply auxiliary information, and the fact that relatively few utilities do most of the electric power business, the best general design is established. Perhaps, however, in the future, regression estimation may be used in place of ratio estimation. At this time, it seems reasonable to implement the Keyfitz method, modified to include a certainty stratum, for sales estimation, and a similarly modified double ratio estimate for prices.

2. PROPOSED SAMPLING PLAN:

Goal -

The purpose of this plan is to provide sales and price estimates at the State level and higher, with estimated State level cv's under 5 percent, while not increasing respondent burden. Because there are several different classes of service (i.e., "sectors") to consider, this is a multivariate problem.

Sample Design -

Thus, for 1990, a new sampling methodology is being implemented. The unit of response is no longer at the utility level, but will be on a State-utility basis. The largest utilities in each State were selected for a certainty stratum. The stratified random sample of smaller State-utility units is taken from the remainder of the universe which consists of approximately 3000 utilities, most of which do business in only one State, and are therefore called "single-Staters." In the new methodology, there is slightly less coverage by the certainty stratum, and more coverage by the random strata. Approximately 70 percent of total U.S. retail electricity sales is to be covered by approximately 200 State-utility units selected with certainty. At least two random strata are found in each State. Thus, this provides for over 100 random strata in the U.S., with over 200 additional units joining the certainties to form the sample. The total sample size was increased by 1) a nominal decrease in average coverage by the certainties, 2) not requiring all State parts from a large utility which has some State parts in which it does little business, and 3) elimination of some information requirements on the form, such as the number of customers and wholesale data. All of the above reduced calculated burden, so that more stratified random sampling could be done. Therefore, State-level sampling is now possible.

The universe from which the Form EIA-826 sample is taken, the Form EIA-861 data, is used for auxiliary information, thus increasing the accuracy of our sales and price estimates. With the new sampling methodology and estimation procedures, the EIA will be able to publish State, Census division, and national level estimates of sales, sales cv's, prices, and price cv's. Similar revenue estimators are available. Unfortunately, formulas for coefficients of variation (cv's) at the State level will be "rough" approximations, as only large sample approximations of cv's using auxiliary data are available.

Coverage -

Within each State, coverage, based on the 1987 Form EIA-861, is usually at least 70 percent for each sector. There is at least 2/3s coverage in the residential, commercial, and industrial sectors for both sales and revenue, 1/2 coverage for other sales and revenue, and at least 2/3s coverage of total sales. The remainder of the Form EIA-861 respondents could then be divided into (at least) two strata per State for stratified random sampling. There are some exceptions to the amount of coverage by the certainty strata. In Minnesota, for example, residential sales and revenue coverages could not reach 2/3s without fifteen respondents in the certainty stratum. However, if the coverage by certainties requirement is relaxed, to 60 percent, then the number of respondents needed from this stratum is reduced to eight. The reason for this problem is that in Minnesota, residential (and to a lesser extent, commercial) sales (and revenue) are distributed widely. That is, a large number of utilities in that State have substantial residential sales. The best action to be taken under such circumstances may be to increase the number of strata for random selection, and in this case four were used. (See "Suggested rules for establishing strata for random selection" found later in this paper.) Two observations were drawn from each such stratum, for a total selection of 16 units from Minnesota. This is almost twice the average sample size per State, but such exceptions must be expected.

Certainty selection algorithm -

This program was written to ensure, unless overriden, that the minimum coverages described above for sales and for revenue in the various class of retail service sectors will be present in the certainty stratum. Each time sales or revenue is examined for a new sector, a check is first made to see if units already selected have completed the requirement in this new area as well. If not, only the difference must be taken into account.

A set of records containing revenue and sales data for the four independent class of service sectors, plus a fifth "total" sector, with each record providing data for a given utility in the State of interest (i.e., one record is found per unit), is sorted by total sector sales and used as input. This file is constituted of auxiliary census data for a given State. This file is then studied using the following steps:

1) Starting with the largest total sales record,

pick consecutively ranked records until 2/3s of the total sales in the auxiliary data file are "covered."

2) Next consider residential revenue, followed by residential sales, commercial revenue, commercial sales, industrial revenue, industrial sales, "other" revenue, and "other" sales. Examine the records for the units already chosen as certainties based on total sales, to see what coverage is already accomplished for residential revenue.
3) If residential revenue already has 2/3s coverage, go on to residential sales, etc.
4) Otherwise, starting with the record not selected thus far, which has largest total sales, and proceding through the file, determine whether any one record could be added, which would complete the required coverage for the category.
5) If this fails, return to step 4, but look for records which will supply at least 1/4 of the required coverage remaining.
6) If any records are selected through step 5, the search is continued until either the coverage requirement for this category is completed, or up to three such records are found, and a final record can be chosen which must only be large enough in the category of interest to complete the required coverage.
7) Finally, if the required coverage is not completed, the program returns to the record with largest total sales which has not yet been selected, and selects each such record, in order of total sales until the requirement for the category under study is completed.
8) Go to the next category, see if the coverage requirement (usually 2/3s for all categories, except 1/2 for other revenue, and sales) has been satisfied by units selected previously, and if not, go to step 3, and repeat for all categories in turn.

Note that this algorithm is applied for various component sales and revenues against the file sorted only by total sales. Some of the large total sales units will not be included, but there is a tendency to select them. This is helpful in that if two records are each capable of completing a given requirement, it is better to select the record which is most likely to help with coverage in other categories. Also note that exceptions to the above algorithm were based on unusual data distributions.

Suggested rules for establishing strata for random selection -

Strata from which random selection is to occur are decided upon with the goal of reducing the coefficients of variation (cv's) for residential, commercial and industrial sales and revenue, simultaneously, in an approximately optimal fashion. Cochran (1977), notes that such an allocation is population dependent. In our case, there are tradeoffs to be considered between six variables (residential, commercial and industrial, sales and revenue). The possibility of adjustments at the Census division level, for purposes of improving (reducing) cv levels, is discussed in the next subsection.

Only cv's were considered here. (The bias of the ratio estimate is also important, but can be studied in relationship to the cv of the auxiliary variate. With our data, this does not appear to be a problem. The Keyfitz method as applied here is "...without bias...," Keyfitz (1957)). Strata may be redefined in succeeding years on the basis of results of the first year(s) of implementation.

If a unit selected with certainty has merged with another certainty, no adjustment to these procedures will be necessary. If it merges with a noncertainty, weights will have to be adjusted, and similarly, if a randomly drawn unit is involved in a merger.

Only two units per random stratum are drawn so that more strata can be used, and to make use of the Keyfitz method. Strata, from which random selections are to be made, are limited to 200 units to avoid very large weights, and particularly in case there is a lot of variance within such a stratum which may be poorly measured when drawing only two units from each.

Each of the 50 States and the District of Columbia are considered. The remaining potential respondents were examined manually after the certainty selection program had been run. For special cases, such as mentioned earlier in the case of Minnesota, the general goal was still the same. An effort was made to minimize the cv's expected for sales and for revenue in the residential, commercial and industrial sectors, while trying not to greatly exceed an average of eight respondents per State so that respondent burden would not be increased. (Note that cooperatives and municipals that buy electricity from the Tennessee Valley Authority (TVA), whose sales and revenue data are collected by TVA under the TVA Act, are included as certainties and only counted as a small burden to match historical burden apportionment since no additional respondent burden is required.) This system is now being implemented. In the future, EIA contracting support may move toward more complete automation, as in an expert system, to some degree, in the selection of strata boundaries. In addition to computerizing calculations using the formulas given in Appendix A of this paper, the possibility of automation has been discussed for at least part of the strata definition which was done more subjectively here. (Any stratification can be changed in succeeding years if this proves to be necessary.) Rules given below concentrate on cases where only two strata are needed, and the original goals for coverage by certainties have been met. With these restrictions in mind, the contractor has been given the following general rules for determining strata, which would provide at least a more objective basis for deciding upon the detailed construction of the sample:

i) If fewer than eight potential respondents remain after the certainties have been removed from the Form EIA-861 listing for the given State, then stop. I will want to examine this.
ii) If, after the following rules are applied, fewer than three, or more than 100 units ("potential" respondents) are found in one of the two strata into which we are attempting to place the noncertainties, then also stop.

Otherwise:

iii) Calculate Nh*SQRT(Xhij) for candidate strata, where h represents one of the (in this case) two strata which are candidates for random selection, i indicates sales or revenue, and j represents the residential, commercial or industrial sector.
iv) Calculate NhXhij, the stratum total, for each stratum, i and j.
v) For each of the six i,j combinations, calculate Dij = ([N1*SQRT(X1ij) - N2*SQRT(X2ij)]/ N1$QRT(X1ij))*100, to express this as a percent. h=1 is for the "larger" potential respondents, where we probably have N1<N2. Then calculate D=Σ i,j Di,j.
vi) Similarly, calculate D', where square roots are not used.
vii) I made some preliminary suggestions for minimum acceptable D and D' values, and for cvhij if the finite population restriction is dropped.
viii) Let SCC represent the portion of sales due to the certainties (in a given sector) - i.e., SCC= xstc/Xs - and similarly, for revenue. (See Appendix A.) For noncertainties, ESCNC= (Xst-xstc)/Xs, and similarly for revenue. Ideally, SCC+ESCNC would be approximately 1.0, and this would also be true for the revenue equivalent. Calculate ESCNC and the revenue equivalent for the candidate strata for the worst case scenarios. This may be used when alternative stratifications appear approximately equally good by other criteria above.

Results of Length of Initial Run Test applied to the New England Census Division -

The Length of Initial Run (LIR) Test was originally designed as a test for heterogeneity between sets of respondents from different populations that were presumed to be similar. The purpose of the test was to examine the advisability of continuing to group these units, based on limited testing. (See Knaub, et.al. (1983).) Here, this test is used to determine if there is a high degree of homogeneity between strata when calculating Census division level numbers. For example, at the Census division or national level, it may be found that the strata formed appear quite heterogeneous for some variates, but that for others, no one stratum seems to have noticeably larger values than the others. Perhaps a fine adjustment of the strata may be needed. To help reduce cv(s) for the area(s) of interest, one response might be to add another certainty respondent to the sample. Here, only the New England Census Division was studied. No adjustment appeared to be necessary in this case.

The null hypothesis is that all pairs of observations come from the same population. As indicated graphically, following the text of this paper, if 20 observations have been drawn as 10 pairs of observations (i.e., two observations are drawn from each of 10 i.i.d. populations), and the results for a given variable have been ranked, one must determine the number of observations, starting with the largest, that appear before the smaller one from any pair. There is approximately a 0.053 probability of an LIR of 1, a 0.105 probability of an LIR of 2, a 0.149 probability of an LIR of 3, etc. The expected value of the LIR in this case is approximately 4.7. If, however, 9 pairs of observations are drawn from a N(x,1) and 1 pair from a N(x+2,1), then the expected value for the LIR is about 2.6. Here, the approximate probabilities of LIRs of 1, 2 and 3 are 0.36, 0.22 and 0.15, respectively. If 5 pairs were drawn from a N(x,1) and 5 pairs were drawn from a N(x+2,1), the expected LIR would be 3.3.

3. PLANNED SALES, SALES CV, PRICE AND PRICE CV ESTIMATORS:

For the Form EIA-826 monthly sample, it is planned that, beginning in January 1990, ratio estimation will be used at the State level, with auxiliary data coming from the Form EIA-861 annual census for calendar year 1987, with a possible update to 1988. For sales in each State, for each end-use sector, use is made of the Keyfitz formulation, modified to include a certainty stratum from which all units are observed as opposed to the random selection of two units each required in the Keyfitz formulation for each of the other strata. This is accomplished as a combined estimate. All sales aggregate values (i.e., for the "total sector," and Census division and national levels) are added using separate estimation. Thus, confusion is avoided for nonstatistician users of the data who are expecting data to be additive. Price

estimation is accomplished using a similarly modified, combined double ratio estimator, with separate estimation used to accumulate these results to the aggregated levels. Estimators are given in detail in Appendix A.

As another alternative, note that ratio estimation could be used on the stratified random portion of the sample only, and the certainties treated separately. This would alter Appendix A formulas, treating the certainties as a separate entity (thus, requiring more use of the separate method). This alternative, however, may not be the better model for our data, as random sampling is limited such that a completely combined estimate (at the State level) may be advisable. Past experience shows that when a ratio-like estimate of State-level sales (using primarily certainties) was employed, results were very good when later compared to census data for that period.

National and even Census division level cv estimates will be more reliable, but the State-level cv estimates will be very rough due to the small sample sizes. For a given State and sector, these estimates may have greater variability from month to month than the true cv's do, but even more volatility will probably be found when samples are redrawn (perhaps every two or three years). This can be studied in the long term once these data are available for a number of samples drawn. In the interim, as long as we realize that cv estimates for sales at the State level are very rough estimates, price estimation is good due to further ratioing, and improvement in accuracy of the cv estimates is found as States are aggregated, these results can be useful to the decision maker who wants to know how much sampling error might be reasonably expected.

4. CURRENT AND PROPOSED FUTURE EVALUATIONS:

Test data -

Monthly sales and revenue data have never been censused. It has been suggested, however, that annual census data from 1987 be used as auxiliary data in our test set, just as it may be for our 1990 sampling, but in place of 1990 monthly data, we will use the corresponding data from the 1988 census. Therefore, we will be able to compare the results from our estimators to the complete census for 1988, as if it were the monthly data collected from all utilities. cv estimator performance, as well as the price and sales themselves can be examined this way. This does not test for everything, but does do a lot. Also, the contractor can institute a more complete version of the software since we will be dealing with all annual census (Form EIA-861) codes, and we will now use those same codes (plus a State indicator) for 1990 monthly data. (See Appendix B for test data results.)

Although strata were set up based on the 1987 census of utilities, 1988 census data may actually be used as auxiliary data when this is first used for 1990 monthly estimations. Few births, deaths or mergers are indicated thus far. Data may be studied to see if any anomalies appear when running the inverse of the above procedure. That is, the 1988 census can be used as auxiliary data, and the sample drawn from the 1987 census. Results may then be compared to the 1987 full census results. Whether 1987 or 1988 data are used as auxiliary data for the actual 1990 estimations may depend greatly upon which data set is considered to have the least nonsampling error.

Ratio vs Regression Estimation -

This may be the subject of a very long term study of our data, which could be done most easily and of greatest value, using the data that will be collected under the new sample design starting in 1990. Some simulation from Form EIA-861 data may be of supplemental help; however, a study will be more complete with monthly data drawn by State. Further, from Cochran (1977), page 197, "In small samples on natural populations the regression estimate appears disappointing in performance. In eight natural populations of the type in which the ratio estimate has been used, Rao (1969) found in a Monte Carlo study that the average of the ratios..." of the mean square error of the linear regression estimate to the mean square error of the ratio estimate was 1.15 for a sample of size "... n=12, 1.36 for n=8, and 1.51 for n=6." Bias is of the same order, although it seems easier to quantify bias from the ratio estimate; variance is almost the same in large sample approximations for both the ratio and regression estimates, although such an estimate of the variance of the ratio estimate is slightly larger, or at least not less than that for the regression estimate; and the model for the ratio estimate is simpler, which I feel may have contributed to Rao's results above - i.e., a more sophisticated model may not be justified by the data. All of this indicates that there are reasons for and against use of each of these estimators, and it may not be clear until after a long and major study, which one may be slightly better for our data. I have opted not to use regression at this time, as I feel ratio estimation is satisfactory for these data. The simplest model may be all that is justified. Additionally, implementation is quite practical and can be accomplished for 1990 data.

5. SUMMARY OF PLANNED ESTIMATION:

In the case of the Form EIA-826 monthly sample, beginning in January 1990, ratio estimation will be used at the State level, with auxiliary data coming from the Form EIA-861 annual census for calendar year 1987 (or 1988).

For sales in each State, for ordinary end-use sectors, use is made of the Keyfitz formulation, modified to include a certainty stratum from which all units are observed as opposed to the random selection of two units each required in the Keyfitz formulation for each of the other strata. This is accomplished as a combined estimate. All sales aggregate values (i.e., for the "total sector;" and Census division and national levels) are added using separate estimation. Price estimation is accomplished in a similar manner, using modified double ratio estimation.

Also, remembering that this is done at a State level, and States vary greatly in size and distribution of units, a hypothesis test specialized to study samples of two observations each, has been used in a pilot study of the New England Census Division. One recommendation from that study could have been to increase the number of certainty units in one of the New England States, but that did not appear to be necessary in this case.

The overall goal is to provide as much and as high quality data as possible, without increasing calculated respondent burden. (This design has the additional advantage of spreading burden more evenly than in the past. Most of this advantage, however, is actually due to a reduction in the number of data elements required.) CVs are a measure of the adequacy of a sample design, and sample sizes. (The bias of the ratio estimate is also important, but can be studied in relationship to the cv of the auxiliary variate. With our data, this does not appear to be a problem. The Keyfitz method as applied here is "...without bias...," Keyfitz (1957)). Strata may be redefined in succeeding years on the basis of results of the first year(s) of implementation.

If fewer than eight potential respondents remain after the certainties have been removed from the Form EIA-861 listing for the given State, then the entire State may be censused. If fewer than three, or more than 100 units ("potential" respondents) are found in one of the two strata into which we are attempting to place the noncertainties, then, in the former case, one or more units, depending upon distribution, must be shifted to another stratum (or strata), and in the latter case, more than two strata for noncertainties may be justified, unless they are very close to being homogeneous.

Further, I tried to arrange that at least one stratum in each State from which random sampling is to be done has no zero entries, or no more than one, for any one data element, except perhaps "other" sales and/or revenue, on the F861 auxiliary data file. This helps to insure the influence of smaller units on the estimates. The combined ratio estimate is used at the State level, however, to ensure that a poor estimate of an individual stratum ratio can not have a devastating effect. Also, it eliminates the possibility of division by zero, a special case of a poor estimate of an individual stratum ratio.

As long as approximately 2/3s of the sampling errors are less than the corresponding cv estimates, we may assume that the cv estimates are reasonable, and statistical bias is negligible. This may be a good topic for hypothesis testing, and/or regression analysis. I am investigating this for a possible paper in the future.

Note that the monthly data for 1990 will probably yield higher cv estimates at the State level than the 1988 test data.

6. FINAL REMARKS:

To summarize the above -

In the case of the Form EIA-826 monthly sample, beginning in January 1990, ratio estimation will be used at the State level, with auxiliary data coming from the Form EIA-861 annual census for calendar year 1987. For sales in each State, for ordinary end-use sectors, use is made of the Keyfitz formulation, modified to include a certainty stratum from which all units are observed as opposed to the random selection of two units each required in the Keyfitz formulation for each of the other strata. This is accomplished as a combined estimate. All sales aggregate values (i.e., for the "total sector;" and Census division and national levels) are added using separate estimation to avoid confusing nonstatistician users of the data. Price estimation is accomplished in a similar manner, using modified double ratio estimation.

At the State level, this is a very rough approximation, as it is a "large sample approximation." In this survey, it is suspected that nonsampling error may often exceed sampling error. Therefore, a better estimate of the cv would not be extremely useful in that we have even less information about the nonsampling error. (Edits help with nonsampling error, but seldom are enough resources committed to use them in a nearly optimal fashion. Knaub (1989) may be used to control for "flagging" records which are not in error, and not "flagging" records which are in error, but to know the overall magnitude of the differences in published results caused by these errors would require even more resources. Even with unlimited resources, certain practical limitations will always exist.

Other electric power surveys which may benefit from sampling -

The EIA has a monthly census survey of all utilities in the United States that generate electricity. To reduce respondent burden, and perhaps reduce

nonsampling error in this monthly process, it has been suggested that sampling be done monthly, and an annual supplemental survey used to complete an annual census. The same general design would seem appropriate in that case. There would be an annual census for auxiliary information, and the data are characterized by a relatively few large utilities that dominate the industry. This survey may be a better candidate for regression estimation. A complication arises in that data of interest is categorized by State and fuel type used for generation, records are by fuel type, but data collection would be by plant. Therefore, if a given plant's data is collected to satisfy a requirement for a certainty stratum, data for other fuel types will sometimes be included. This would have to be added to the certainty strata for the other fuel types, and weights for the noncertainty strata adjusted accordingly.

Another EIA survey is a new data collection effort for nonutility generators, such as a farm that generates its own electricity and sells the excess to an electric utility. Currently, only the largest of these nonutility generators ("cogenerators") are surveyed. Perhaps we could use the current data collection effort as auxiliary data for a sample, and with the resources saved, we could use another sample design, perhaps PPS sampling, to estimate generation among the remaining smaller cogenerators.

General notes -

Although sales variance and the variance of the ratio from the Keyfitz method are different, cvs for sales and the ratio are the same formula.

It should be noted that with auxiliary data for ratio (sales), and double ratio (price) estimation, when there is positive correlation between auxiliary (annual census) and monthly (sample) data, cvs are reduced (estimation is improved) because we have additional useful information, beyond the monthly sample itself. Thus, sample size is effectively increased. If, however, correlation is negative, cv's will be increased. There are two "components" to cv's: that part between sampling units, and what is attributable to the relationship between the sample and auxiliary census data.

With ratio estimation, additional, correlated data are used to advantage, and we are able to calculate cv estimates for sales, which is a great advantage over the currently used method which does not. Also, calculation of estimates of sales cv's, price, and price cv's at levels lower than the national level, may be very useful, particularly at the Census division level, as estimates become less accurate with smaller sample sizes at the State level. (Bias, as mentioned earlier, does not appear to be a great problem.)

Note that as long as we can calculate estimates of cv's, we can estimate the cv of the difference between two monthly estimates. Thus, any objections to calculating monthly price data by State (anticipating misuse) are not well founded. Correlation between sales and revenue makes price the better candidate for publication, if only either sales or price should be published.

When estimation is not helped as much by the auxiliary data, estimates of cvs are designed to reflect this.

Acknowledgements

The author thanks Edgar Betancurt for programming support, and Renee Miller, William Kelly, Dean Fennell and other analysts for their interest and/or suggestions, even when we do not always agree on every approach.

References:

Cochran, W. G., (1977), Sampling Techniques, John Wiley & Sons, New York.

Keyfitz, N., (1957), "Estimates of Sampling Variance Where Two Units Are Selected From Each Stratum," Journal of the American Statistical Association, 52, 503-510.

Kish, L. (1965), Survey Sampling, John Wiley & Sons, New York.

Knaub, J. R., Jr., et.al., (1983), "Analyzing n Samples of 2 Observations Each," in Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research, Development and Testing, Research Triangle Park, NC: U.S. Army Research Office, 23-113.

Knaub, J. R., Jr., (1985), "Nonparametric Median Estimation," in Proceedings of the Thirtieth Conference on the Design of Experiments in Army Research, Development and Testing, Research Triangle Park, NC: U.S. Army Research Office, 199-211.

Knaub, J. R., Jr., (1989), "Fellegi-Sunter Record Linkage Theory As Compared to Hypothesis Testing," to appear in the Proceedings of the 21st Symposium on the Interface (INTERFACE '89).

Addendum to references:

Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors," Bulletin of the International Statistical Institute, 47(3), 495-514.

Rao, J. N. K., (1957), "Double Ratio Estimate in Forest Surveys," Journal of Indian Social and Agricultural Statistics, 9, 191-204.

Rao, J. N. K., (1958), "Estimation of the Ratio in Forest Surveys," Indian Forester, 84, 153-161.

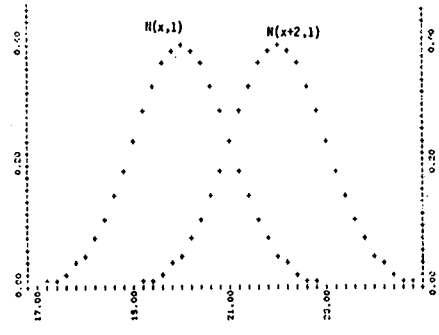
Rao, J. N. K., and Pereira, N. P., (1968), "On Double Ratio Estimators," Sankhya, A30, 83-90.

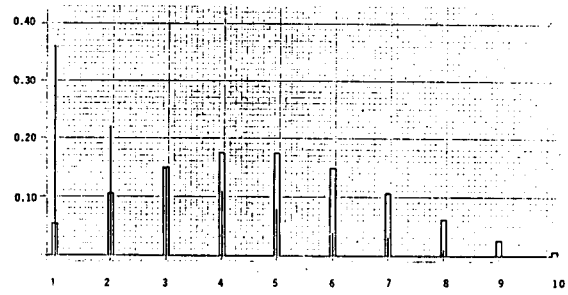
In the New England Census Division, there are six States. One, Rhode Island, has only six electric utilities and is therefore censused. The remaining five States were assigned the minimum of two strata each from which random sampling is to occur. Therefore, there are 10 strata for random sampling in the New England Census Division. Random samples were taken from the auxiliary data to see what LIRs were generated. In all, 10 samples of such data were drawn, and results are given in the table below for all 10 replications.

LIR for the Largest Responses on the Auxiliary Data

Replication	1	2	3	4	5	6	7	8	9	10	Mean
Residential revenue	1	2	3	1	1	3	1	1	2	5	2.0
Residential sales	1	1	3	1	1	4	1	1	2	5	2.0
Commercial revenue	3	4	2	3	4	1	4	3	4	5	3.3
Commercial sales	3	3	2	2	4	1	4	3	4	5	3.1
Industrial revenue	2	2	1	2	2	4	2	2	1	4	2.2
Industrial sales	2	3	3	1	1	4	2	2	2	4	2.4
Other revenue	1	2	2	1	4	2	1	4	4	2	2.3
Other sales	1	2	2	1	4	3	1	4	4	3	2.5

Using Knaub (1985), we see that these 10 replications will likely yield a good estimate of the median of the LIR in each case. The sample means shown in the table indicate that results are not as likely to be improved as much in the commercial sector in the New England Census Division by the stratification chosen as they are in the other sectors. A quick review of the results of my program to select certainties, however, shows that about 90 percent of all commercial sales and revenue for this Census division are covered by the certainties. Thus, there appears to be no sufficient reason to modify the design.





H_0: 10 pairs of observations drawn from the same distribution

H_1: 9 pairs from N(x,1) and 1 pair from N(x+2,1)

Appendix A

Estimators

$\hat{Y}_s$ = sales estimate

$$= \left[\frac{y_{stc} + \sum_{h}^{N} \frac{h}{2} (y_{shl} + y_{sh2})}{x_{stc} + \sum_{h}^{N} \frac{h}{2} (x_{shl} + x_{sh2})} \right] X_s , \qquad (1)$$

where N_h = number of Form EIA-861 (auxiliary data) respondents in stratum h,

y_{shl} and y_{sh2} are the Form EIA-826 sales respondents randomly selected from stratum h,

y_{stc} is the total of the Form EIA-826 sales respondents in the certainty stratum,

x_{shl} and x_{sh2} are the Form EIA-861 sales respondents in stratum h which correspond to y_{shl} and y_{sh2},

x_{stc} is the total of the form EIA-861 sales respondents that correspond to members of the Form EIA-826 certainty stratum,

and X_s is the Form EIA-861 total for all sales in the given end-use sector and State.

For the "total" sector in each State, and for the Census division and National levels,

$$\hat{Y}_s(\text{aggregated}) = \sum_i \hat{Y}_{si} , \text{ where } \hat{Y}_{si} \text{ is } \hat{Y}_s \text{ above for "lower" level sector and/or State level i.} \qquad (2)$$

Sales coefficient of variation (cv) estimator:

$$\hat{cv}_s = \left| \left[\sum_h (1-f_h) \left(\frac{dy'_{sh}}{\hat{y}_{st}} - \frac{dx'_{sh}}{\hat{x}_{st}} \right)^2 \right]^{1/2} \right| , \qquad (3)$$

where $f_h = \frac{n_h}{N_h} = \frac{2}{N_h}$,

$$\hat{y}_{st} = y_{stc} + \sum_{h}^{N} \frac{h}{2} (y_{shl} + y_{sh2}),$$

$$\hat{x}_{st} = x_{stc} + \sum_{h}^{N} \frac{h}{2} (x_{shl} + x_{sh2}),$$

$dy'_{sh} = (N_h/2)(y_{shl} - y_{sh2})$, and

$dx'_{sh} = (N_h/2)(x_{shl} - x_{sh2})$.

$\hat{y}_{st}$ and $\hat{x}_{st}$ were chosen as denominators in the above cv estimation formula because the corresponding numerators would likely be more highly correlated with $\hat{y}_{st}$ and $\hat{x}_{st}$ than with Ys and Xs, respectively. Therefore, if, for example, x_{shl} is a fair representative of 861 sales in stratum h, but x_{sh2} is one of the largest sales values in stratum h, then $\hat{x}_{st}$ is also likely to be a high estimate, and the overall effect may be to reduce the error in $dx'_{sh}/\hat{x}_{st}$.

$$\hat{cv}_s(\text{aggregated}) = \left| \left(\sum_i (\hat{cv}_{si}^2 \hat{Y}_{si}^2) / \hat{Y}_s^2(\text{aggregated}) \right)^{1/2} \right| . \qquad (4)$$

$\hat{P}$ = price estimate

$$= \left[\left(\frac{y_{rtc} + \sum_{h}^{N} \frac{h}{2} (y_{rhl} + y_{rh2})}{y_{stc} + \sum_{h}^{N} \frac{h}{2} (y_{shl} + y_{sh2})} \right) \Big/ \left(\frac{x_{rtc} + \sum_{h}^{N} \frac{h}{2} (x_{rhl} + x_{rh2})}{x_{stc} + \sum_{h}^{N} \frac{h}{2} (x_{shl} + x_{sh2})} \right) \right] P_A \qquad (5)$$

$= [(\hat{y}_{rt}/\hat{y}_{st})/(\hat{x}_{rt}/\hat{x}_{st})](X_r/X_s)$, where $P_A = X_r/X_s$, and X_s is as defined earlier and X_r is the revenue equivalent.

The subscript "r" is used to designate "revenue" in place of sales. Also, P_A is the value obtained for price using the entire set of Form EIA-861 auxiliary data, dividing total revenue (X_r) by total sales (X_s), using only the auxiliary data.

$$\hat{P}(\text{aggregated}) = \frac{\sum_i \hat{P}_i \hat{Y}_{si}}{\hat{Y}_s(\text{aggregated})} , \qquad (6)$$

where $\hat{P}_i$, like $\hat{Y}_{si}$, is for lower level sector and/or State level i.

The price cv is estimated by $\hat{cv}_p$, where, from Cochran's third edition of Sampling Techniques, pages 183-184, the following is derived:

$$\hat{cv}_p = \left| (\hat{cv}_{826}^2 + \hat{cv}_{861}^2 - 2\hat{cv}_{826\text{-}861})^{1/2} \right|$$

where,

$$\hat{cv}_{826}^2 = \sum_h (1-f_h) \left(\frac{dy'_{rh}}{\hat{y}_{rt}} - \frac{dy'_{sh}}{\hat{y}_{st}} \right)^2$$

$$\hat{cv}_{861}^2 = \sum_h (1-f_h) \left(\frac{dx'_{rh}}{\hat{x}_{rt}} - \frac{dx'_{sh}}{\hat{x}_{st}} \right)^2$$

$$\hat{cv}_{826\text{-}861} = \sum_h (1-f_h) \left(\frac{dy'_{rh}}{\hat{y}_{rt}} - \frac{dy'_{sh}}{\hat{y}_{st}} \right) \left(\frac{dx'_{rh}}{\hat{x}_{rt}} - \frac{dx'_{sh}}{\hat{x}_{st}} \right)$$

f_h, $\hat{y}_{st}$, $\hat{x}_{st}$, dy'_{sh} and dx'_{sh} are as defined earlier.

$\hat{y}_{rt}$, $\hat{x}_{rt}$, dy'_{rh} and dx'_{rh} are revenue equivalents.

Note that $\hat{y}_{st}$, $\hat{x}_{st}$, $\hat{y}_{rt}$ and $\hat{x}_{rt}$ are all found in the formula for $\hat{P}$ as parts of the double ratio.

From Kish, Survey Sampling, page 503, this can be written in the algebraically equivalent form

$$\hat{cv}_p = \left| \left\{ \sum_h (1-f_h) \left[\left(\frac{dy'_{rh}}{\hat{y}_{rt}} - \frac{dy'_{sh}}{\hat{y}_{st}} \right) - \left(\frac{dx'_{rh}}{\hat{x}_{rt}} - \frac{dx'_{sh}}{\hat{x}_{st}} \right) \right]^2 \right\}^{1/2} \right| \qquad (7)$$

Kish mentions a form that requires fewer divisions, but the formula given above is easy enough to implement and makes use of terms already calculated.

Now, note that since

$$\hat{cv}_s = \text{estimate of sales cv} = \left| \left(\sum_h (1-f_h) \left(\frac{dy'_{sh}}{\hat{y}_{st}} - \frac{dx'_{sh}}{\hat{x}_{st}} \right)^2 \right)^{1/2} \right| \quad (3 \text{ repeated})$$

it follows that

$$\hat{cv}_r = \text{estimate of revenue cv} = \left| \left(\sum_h (1-f_h) \left(\frac{dy'_{rh}}{\hat{y}_{rt}} - \frac{dx'_{rh}}{\hat{x}_{rt}} \right)^2 \right)^{1/2} \right|. \quad (8)$$

Letting $\hat{cv}_s = \left| \left(\sum_h (1-f_h) \hat{scv}_h^2 \right)^{1/2} \right|$ and $\hat{cv}_r = \left| \left(\sum_h (1-f_h) \hat{rcv}_h^2 \right)^{1/2} \right|$, and rearranging some terms in $\hat{cv}_p$, it follows that

$$\hat{cv}_p = \left| \left(\sum_h (1-f_h) (\hat{scv}_h - \hat{rcv}_h)^2 \right)^{1/2} \right|. \quad (9)$$

Note that dy'rh and dx'rh are the revenue equivalents of dy'sh and dx'sh, respectively.

Finally, $\hat{cv}_p$ (aggregated) =

$$\left| \left(\sum_i \left[\hat{cv}_{pi}^2 \hat{P}_i^2 \left(\frac{\hat{Y}_{si}^2}{\hat{Y}_{s(\text{aggregated})}^2} \right) \right] / \hat{P}^2(\text{aggregated}) \right)^{1/2} \right| \quad (10)$$

Here, $\hat{cv}_{pi}^2 \hat{P}_i^2$ is the variance estimate of $\hat{P}_i$, and since the square of a constant multiplied by the variance of a random variable is the variance of the product of that constant and the random variable, if $\hat{Y}_{si} / \hat{Y}_{s}(\text{aggregated})$ is treated as a constant, then what is inside the square brackets above is the variance of the ith term of $\hat{P}$(aggregated) in equation (6). Assuming independence for all i, this says that the summation in equation (10) provides the variance of $\hat{P}$(aggregated).

To restate equation (6) more appropriately here,

$$\sum_i \left| \left(\frac{\hat{Y}_{si}}{\hat{Y}_{s(\text{aggregated})}} \right) \hat{P}_i \right| = \hat{P}(\text{aggregated})$$

Appendix B

Preliminary Test Results

Census division level cv estimates for all classes of retail service (sectors), expressed in decimal fractions of percents:

Census division	residential	commercial	industrial	other	total
	sales; price	sales; price	sales; price	sales; price	sales; price
New England	0.06; 0.08	0.14; 0.51	0.14; 0.27	1.41; 4.82	0.07; 0.22
Middle Atlantic	0.14; 0.11	0.04; 0.12	0.04; 0.32	0.04; 0.09	0.05; 0.09
East North Central	0.07; 0.08	0.16; 0.14	0.12; 0.15	0.80; 0.83	0.07; 0.10
West North Central	0.32; 1.07	0.70; 0.35	0.61; 0.57	2.09; 1.14	0.30; 0.51
South Atlantic	0.31; 0.29	0.34; 0.15	0.31; 0.25	3.20; 0.41	0.21; 0.15
East South Central	0.15; 0.24	0.08; 0.19	0.03; 0.07	0.44; 0.29	0.06; 0.11
West South Central	0.24; 0.31	0.39; 0.15	0.30; 0.14	2.58; 0.76	0.19; 0.14
Mountain	0.15; 0.14	0.10; 0.11	0.20; 0.14	1.66; 3.43	0.11; 0.15
Pacific Contiguous	0.13; 0.16	0.22; 0.10	0.19; 0.24	1.01; 0.44	0.11; 0.09
Pacific Noncontiguous	0.09; 0.17	1.17; 1.06	0.75; 1.06	4.07; 3.81	0.46; 0.48

Note that the expectation of finding noticeably lower cv estimates for price than for sales was not, in general, fulfilled. However, perhaps this could be the result of different biases in the estimators.

	residential	commercial	industrial	other	"total"
	sales; price	sales; price	sales; price	sales; price	sales; price
national level	0.08; 0.12	0.11; 0.06	0.08; 0.08	0.73; 0.34	0.06; 0.06

NEW ENGLAND CENSUS DIVISION

Total sector - sales in gigawatthours, cv's in percents

State	Sales estimate	Sales cv estimate	EIA-861 Census result	Absolute Percent Difference
Connecticut	26916	0.06	26923	0.03
Maine	11219	0.12	11264	0.40
Massachusetts	44677	0.14	44727	0.11
New Hampshire	8855	0.30	8848	0.08
* Rhode Island	6220	N/A	6220	N/A
Vermont	4456	0.31	4416	0.91

"Total" sector - price in cents/kilowatthour, cv's in percents

State	Price estimate	Price cv estimate	EIA-861 Census result	Absolute Percent Difference
Connecticut	8.39	0.06	8.38	0.12
Maine	6.67	0.08	6.70	0.45
Massachusetts	7.67	0.51	7.80	1.67
New Hampshire	8.28	0.20	8.27	0.12
* Rhode Island	7.94	N/A	7.94	N/A
Vermont	8.07	0.48	8.10	0.37

Note that because large sample approximations are being used, the State level cv estimates published will be rounded to tenths of percents.

* Note that Rhode Island, with only six utilities, will be censused each month for sales and revenue data.