Projected Variance for the Model-Based Classical Ratio Estimator: Estimating Sample Size Requirements

James R. Knaub, Jr.
U.S. Energy Information Administration, Forrestal Bldg., Washington DC 20585[1]

Abstract:

Here we explore planning for the allocation of resources for use in obtaining official statistics through model-based estimation. Concentration is on the model-based variance for the classical ratio estimator (CRE). This has application to quasi-cutoff sampling (simply cutoff sampling when there is only one attribute), balanced sampling, econometric applications, and perhaps others. Multiple regression for a given attribute can occasionally be important, but is only considered briefly here. Nonsampling error always has an impact. Allocation of resources to given strata should be considered as well. Here, however, we explore the projected variance for a given attribute in a given stratum, for resource planning at that base level. Typically one may consider the volume coverage for an attribute of interest, or related size data, say regressor data, to be important, but standard errors for estimated totals are needed to judge the adequacy of a sample. Thus the focus here is on a 'formula' for estimating sampling requirements for a model-based CRE, analogous to estimating the number of observations needed for simple random sampling. Both balanced sampling and quasi-cutoff/cutoff sampling are considered.

Key Words: Classical Ratio Estimator, Volume Coverage, Measure of Size, Model-Based Estimation, Official Statistics, Resource Allocation Planning, Sample Size Requirements, Weighted Least Squares Regression

1. Introduction

There are surveys conducted at the US Energy Information Administration (EIA) for which a cutoff, or quasi-cutoff establishment survey is conducted. That is, there is a measure of size that is established, and the largest potential respondents are to be collected. Brewer(1963) called this a “partial collection,” for which he noted prediction could be used to estimate for out-of-sample cases, and that a cutoff sample provides the lowest estimated variance of the total. However, because a respondent generally responds for more than one attribute/variable, some of the smaller potential responses for a given attribute are also collected. This is quasi-cutoff sampling. For the key attributes one may obtain nearly a true cutoff sample, but for other attributes, that is not so much the case. (See Douglas(2007).) Often it is not possible to obtain high quality data from the smallest respondents on a frequent basis, but obtaining some smaller responses from reliable respondents in this manner may help guard against model-failure. Many design-based (randomization) oriented statisticians are concerned with the fact that many smaller

[1] The views expressed are those of the author, and are not official US Energy Information Administration (EIA) positions unless claimed to be in an official US Government document.

members of a population may have no chance of selection. However, placing the missing data where it does the least harm, near the origin, especially where regression through the origin is best, and it generally is best (consider Brewer(2002), page 110), appears not so harmful in the survey data experience of this author. (See Knaub(2010).)

However, because it may be important to collect as much volume with as few observations as possible, there is a tendency for some to put emphasis on this volume coverage only, without considering the variance inherent in the regression relationship between the attribute of interest, and the regressor or regressors. Note that in official statistics, where we have frequent sample surveys, and less frequent census surveys, there is generally one very good regressor, often the same data element from a previous census, but not always. Occasionally there is another good regressor. (See Knaub(2003).) Often there is not, and a second regressor should not be used unless clearly helpful. (See Brewer(2002), page 109.) One should then consider the inherent variance in this regression relationship, and not just the portion of the estimated totals that are 'covered' by the actually observed data. In some cases, a 60-percent coverage may be quite adequate. In other cases a 90-percent coverage may not be very good, and then one may have a greater problem with nonsampling error as the sample size becomes unmanageable on a frequent basis. Results here will not fully account for this. Note that the relative standard error (RSE) for a census is zero, no matter how much nonsampling error there may be. (For further information, see Knaub(2004).)

Nonsampling error can be found both empirically and philosophically discussed, especially in the context of cutoff and quasi-cutoff sampling with model-based estimation, in the following combination of papers**:** Knaub(2001), Knaub(2010), Knaub(1999) pages 8 and 9, Knaub(2004), Knaub(2007c), and Knaub(2002). Revision "error" tables are found through the EIA website for various surveys. They show the magnitude of revisions made to the data as a proxy for the level of nonsampling error.

So, it is important to at least first see how a given volume coverage may translate into a relative standard error or confidence limits for a given estimated total, in addition to considering other problems, such as nonsampling error and more complicated sampling schemes. Knaub(2012b) develops the necessary relationships for this, to see what coverage may be helpful for future data collections, based on inherent variance in similar past surveys. This is analogous to the estimation of a sample size requirement for a simple random sample, as found in Cochran(1977), page 77. Instead of directly finding a sample size requirement, however, Knaub(2012b) finds a coverage requirement. To obtain that coverage requirement with the smallest sample size, a cutoff sample is required. Several examples regarding coverage requirements are included in Knaub(2012b), including an addendum. In this paper we will look at just one more example, and see how this relates a given quasi-cutoff sample size to what could be obtained with the same sample size, but with a strict cutoff, and also compared to a 'balanced' sample, as described in Brewer(1999), and Royall and Cumberland(1978). Basically, in a balanced sample, the sample is chosen to correspond to a regressor, or size variable, that has the same mean value for that subset associated with the sample as it does for the population as a whole. The measure of size can be a regressor (see Table 1, page 2 in Knaub(2012a)), or a linear combination of regressors in multiple regression (page 10, Knaub(2012a). Several options for use of regression prediction are laid out in Knaub(2011b), and a very flexible multiple regression methodology, with small area estimation, is found in Knaub(1999, 2001). Regression through the origin with two regressors is examined in detail in Knaub(1996). Emphasis here, however, remains on the CRE for one regressor. For an excellent description of both design-based and model-based classical ratio estimation, see Lohr(2010), Chapter 4. Note also that although Knaub(2011b) mentions weekly samples in the title, that “Option 1” from page 10 is actually currently being used for a monthly quasi-cutoff sample of natural gas deliveries, with an annual census employed for regressor data.

Note that even if a probability design is to be used, models are often useful in survey planning. See Holmberg(2003), and Brewer(1963).

2. Coverage and Relative Standard Error Estimators

From Knaub(2012b), pages 2 and 3, one can see how heteroscedasticity may be handled for model-based estimation. (For other thoughts on heteroscedasticity, see Knaub(2007a), Knaub(2011c), and Lee(2013).) The classical ratio estimator implicitly assumes a level of heteroscedasticity that is lower than often actually measured for most survey applications, but appears to be fairly robust against commonly occurring nonsampling error conditions. See Knaub(2005). (Also note the caution in Holmberg and Swensson(2001).) The model-based classical ratio estimator (CRE) is quite useful in a number of areas, including econometrics. Thus the estimators developed relating coverage and relative standard error in Knaub(2012b) are specifically written for the CRE. They could be rewritten more generally, from what is in Knaub(2012b), but here we once again concentrate on the CRE, so useful in survey statistics and econometrics. The estimates of totals here also relate to estimates of prices, and any other ratios of totals in that each total must be estimated, and its variance. In addition, covariance between numerator and denominator random variables would be involved. For multiple regression estimators of totals, covariance between estimated regression coefficients is also a concern. For regression through the origin with two regressors, see Knaub(1996). Also, multiple regression may be somewhat simulated by a one regressor model, at least for some planning purposes, by collapsing the regressors into one regressor, a linear combination of them, ideally an estimate of y. (See Knaub(2005), bottom half of page 8, and next to last paragraph on page 9, which leads to Knaub(2010), page 9.) In Knaub(1996), the estimated standard error of the random factors of the estimated residuals, oddly still just commonly referred to as the square root of the “MSE,” occurs in several parts of the relevant estimators. It is the key element to relating the coverage to the relative standard error (RSE) estimates, and here it is written as $\sigma_{e_0}^*$.

Considering weighted least squares regression through the origin, with one regressor, we have the following:

$$y_i = bx_i + e_{0_i} w_i^{-1/2} \qquad \boxed{\text{Eq(1)}}$$

Using $w_i = x_i^{-2\gamma}$, as in Brewer(2002), we have $y_i = bx_i + e_{0_i} x_i^{\gamma}$ where γ is the “coefficient of heteroscedasticity.” When γ = ½, we have the CRE. In this case, the square of $\sigma_{e_0}^*$, or “MSE” is as follows:

$$\sigma_{e_0}^{*2} = \sum_{i=1}^{n} e_{0_i}^2/(n-1) \qquad \boxed{\text{Eq(2)}}$$

In this case, $V_L^*(T^* - T) = \sigma_{e_0}^{*2} \sum_{i=n+1}^{N} x_i + V^*(b)\left[\sum_{i=n+1}^{N} x_i\right]^2$ Eq(3)

This is equivalent to V_1 in Karmel and Jain(1987), page 57. More relevant details are found in Knaub(2009), and Maddala(2001), page 85. Generally, when γ is estimated from the data, it is larger than 0.5, but using this value seems robust against nonsampling error. See Knaub(2005), pages 2 and 3.

Letting "*a*" represent the volume coverage, for a CRE, with one regressor, we have the following**:**

$$a = \sum\nolimits_{i=1}^{n} x_i \Big/ \sum\nolimits_{i=1}^{N} x_i \qquad \text{Eq(4)}$$

This is the inverse of the expansion factor shown in the bottom half of page 1 in Knaub(2012a), and is presented on page 4 of Knaub(2012b). Consider then that the estimated relative standard error (RSE) is the square root of the estimated variance of the total, divided by the estimated total. We see on page 5 of Knaub(2012b), that in terms of the coverage, the portion of the frame for which we observe data, *a*, where an asterisk (*à la* Maddala) represents a weighted least squares regression estimator, gives us the following**:**

$$RSE^* = \sigma_{e_0}^* \sqrt{T_x\left(\frac{1-a}{a}\right)} \Big/ T^* \qquad \text{Eq(5)}$$

Note that the relative standard error, RSE, is generally multiplied by 100% and expressed as a percent. T* is an estimate of the total for the attribute of interest, and $T_x = \sum_{i=1}^{N} x_i$ **.**

Perhaps a better way to write equation 5, so that it more closely resembles what Cochran(1977) has on page 77 for random sampling, would be as follows**:**

$$RSE^* = \left(\sigma_{e_0}^*/T^*\right) \sqrt{T_x\left(\frac{1-a}{a}\right)} \qquad \text{Eq(5a)}$$

Or … $RSE^* = \left(\sigma_{e_0}^*/T^*\right) f(a, T_x)\ or\ \left(\sigma_{e_0}^*/T^*\right) f\left(\sum_{i=1}^{n} x_i , \sum_{i=1}^{N} x_i\right)$

Note that equation (4.5) in Cochran(1977) can be rewritten as $RSE \approx (S/\hat{T}) f(n, N)$.

Rewriting equation 5 as an expression for coverage, on page 15 of Knaub(2012b), we have this**:**

$$a^* = \left\{1 + \left[\left(\frac{(RSE)(T^*)}{\sigma_{e_0}^*}\right)^2 \Big/ T_x\right]\right\}^{-1} \qquad \text{Eq(6)}$$

Examples of the use of the above expressions are found in Knaub(2012b). As noted, here we look at one example, and follow through from the coverage to estimates of sample size, to make the analogy to the estimation of anticipated sample size for a simple random sample in Cochran(1977), page 77, more complete.

3. Anticipating Model-Based CRE Sample Size Requirements

The purpose of equation 6 above is to provide coverage needed when planning a sample, if one can "guess in advance" (*à la* Cochran(1977), bottom of page 77) the necessary input variables. However, once one has decided what coverage (based on testing of archived data, or whatever else is available of relevance) is likely to provide a result with a given level of standard error, that coverage then must be translated into a sample. If only one attribute (variable of interest/question on a survey) is being collected in the current sample, then a strict cutoff sample yields the smallest sample needed. (See Brewer(1963), Royall(1970), and Cochran(1977) page 160.) If too large of a sample is needed on a frequent basis, nonsampling error may become more of a problem. Thus a cutoff sample is useful in avoiding this. Many statisticians object, however, if a portion of the population has no chance of selection. This is addressed in Knaub(2010). Also, see Karmel and Jain(1987), Kirkendall(1992), and Knaub(2007b) regarding positive results without randomization.

Generally there are multiple attributes of interest, and a respondent collected because it has a large size (Knaub(2012a), page 2) for one attribute, will generally report for others as well. This can be problematic for a design-based survey, but additional good data from a reliable source for a smaller response may be helpful in avoiding any problems with model-failure (Knaub(2010)) when using model-based estimation. The sample size needed to achieve a given coverage, as defined above, may be based on a cutoff, and these additional respondents may cause further improvement, if this does not produce too much burden. However, when coverage from these additional responses to a given attribute becomes a large part of overall coverage for that attribute, then the original cutoff can be relaxed (Douglas(2007)). This can be solved using operations research techniques, but trying to be too precise is not justified given the transient nature of the random variables involved. A little trial and error may be sufficient. Such a sample, based on cutoff sampling but with other responses for respondents not large for a given attribute, may be referred to as a quasi-cutoff sample, with respect to a given attribute.

Another sample that may be drawn based on the coverage requirement estimated by equation 6, which may attract less ire from some statisticians steeped in randomization techniques, would be balanced sampling. For a balanced sample, with one regressor, the mean of the x-values associated with the sample should be as near as possible to the mean of the x-values for the entire population. Further, one could even estimate the expected random sample size needed to attain a given coverage, considering the size of each possible respondent, where size for one-regressor modeling would still be the corresponding regressor value x_i. This might best be obtained by simulation. (Various design-based samples could be considered, but this is still with regard to CRE model-based estimation.) A balanced sample, on the other hand, might be somewhat simpler to employ.

Here we compare results for cutoff, quasi-cutoff, and balanced sampling, for given test data sets, for given sample sizes, which result in different coverage levels, and therefore different levels of relative standard error. Thus we use equation 5a to show how these different sampling methods will impact the coverage, and thus the estimated relative standard error.

4. Example: Comparison of these Sampling Methods

Smaller sample sizes are generally desirable when collecting data on a frequent basis, to alleviate respondent burden, cost, and nonsampling error. Small population sizes may often be encountered in official statistics when many such small establishment survey populations are actually part of a very much larger one. Thus sample sizes may be very small. ("Borrowing strength" through small area estimation may sometimes be necessary, as noted in Knaub(2011b) and used in Knaub(1999).) The examples below use test sample data from the EIA-826, "Monthly Electric Utility Sales and Revenue Report with State Distributions," with regressor data for the same data element taken from the EIA-861, "Annual Electric Power Industry Report," currently an annual census, but soon to be a less frequent census. Here, for purposes of this demonstration, we are using "Option 1" from page 10 of Knaub(2011b), a straightforward CRE application. However, small area estimation (page 11 in Knaub(2011b)) may be helpful in some cases, and multiple regression, also discussed briefly in Knaub(2011b) is sometimes a consideration, though not often. (See Brewer(2002), pages 109 and 110 for a caution.)

4.1 Quasi-cutoff Sampling

Consider residential electric sales in Tennessee, by full service providers (*i.e.*, including transmission) falling into two of the ownership categories**:** (A) electric cooperatives, and (B) municipalities.

For (A) we have a test data set with a population of just N = 25, with a quasi-cutoff sample size n = 9. In this example, 25 x-values for the population are from a segment of data collected on the EIA-861 for the year 2011, and the monthly sample selection of 9 y-values is taken from the corresponding part of the EIA-826, for December 2012.

Because the x-values are the measure of size in this one-regressor case (see Table 1, page 2 in Knaub(2012a)), the data used showed that the y-values were collected for the first through eighth, and twenty-fourth largest of the 25 members of the population in this category.

Note that sometimes a y-value will be much larger or much smaller than expected, based on the measure of size. This makes the standard error larger. This can be ameliorated by stratification (see Karmel and Jain(1987), Knaub(2010)), step-functions for the regression weight (Knaub(2009), pages 5 and 6), and sometimes multiple regression (Knaub(1996), Knaub(2003). There are several reasons found in Knaub(2010) for the success of cutoff and quasi-cutoff sampling, as demonstrated in Knaub(2001). The electric sales data shown here have low variance, and good data quality, as can be demonstrated by examining scatterplots with EIA-826 data on the y-axis, and corresponding EIA-861 data on the x-axis. But other data for which equation 6 above where applied in Knaub(2012b), including the addendum, and test data in Knaub(2013), another version of this paper, show a range of levels of variance.

For the test data set (B), N = 61. There are 61 x-values taken from the 2011 EIA-861 census, and n = 11 y-values from the EIA-826 December, quasi-cutoff sample. Using the x-values as measures of size, the sample contained the first through the seventh largest members of the population in this category, and the tenth, eleventh, fourteenth and fifteenth largest of these 61 members.

Some results follow.

4.1.1 For test Case A (cooperatives)

$\sum_{i=1}^{n} y_i =$ 745,345 and $\sum_{i=1}^{n} x_i =$ 9,579,683 in the same units. The slope, b, is the former divided by the latter.

b = 0.0778, s.e.(b) = 0.0016.

Note that as 31/365 = 0.0849, because the month involved had 31 days, this CRE regression coefficient could represent a decline in the volume of residential electric sales in Tennessee. The standard error of the coefficient, b, is small enough to support such a conclusion. Considering seasonality, it does not seem that December should be a particularly low point for residential electric sales. So this result may be of considerable practical interest.

Continuing,

$\sum_{i=1}^{N} x_i =$ 13,657,816 = T_x and $\sum_{i=n+1}^{N} x_i =$ 4,078,133

So here the coverage was a = 9,579,683 / 13,657,816 = 70.1%.

$\left[\sigma_{e_0}^{*}\right]^2$, which one may see referred to as "MSE" (see Knaub(2009), top of page 7), even though it is only based on the random factors of CRE estimated residuals, not the estimated residuals themselves, is here approximately 24.43. So $\sigma_{e_0}^{*} \approx 4.94$.

T* = 1,062,643 and the estimated standard error of this estimated total is 11,917.

RSE* ≈ 1.1%.

4.1.2 For test Case B (municipals)

$\sum_{i=1}^{n} y_i =$ 1,425,661 and $\sum_{i=1}^{n} x_i =$ 19,742,746 in the same units.

b = 0.0722, s.e.(b) = 0.0023.

> Compare this to Case A, full-service cooperatives in Tennessee for residential sales where we had

b = 0.0778, s.e.(b) = 0.0016.

These appear to be close, but perhaps discernible. Both seem low.

Continuing with Case B:

$\sum_{i=1}^{N} x_i = 28{,}712{,}373 = \mathrm{T}_x$ and $\sum_{i=n+1}^{N} x_i = 8{,}969{,}627$

So here the coverage was a = 19,742,746 / 28,712,373 = 68.8%.

$\left[\sigma_{e_0}^{*}\right]^2$ is approximately 103.2. So $\sigma_{e_0}^{*} \approx 10.16$.

T* = 2,073,375 and the estimated standard error of this estimated total is 36,698.

RSE* ≈ 1.8%.

4.2 Strictly Cutoff Sampling

4.2.1 For test Case A (cooperatives)

4.2.1.1 Results for Same Sample Size as for Data Collected

If the cases with the largest n = 9 x-values had been the respondents, then results would have changed as follows:

Using $\sigma_{e_0}^{*} \approx 4.94$, and the new $\sum_{i=1}^{n} x_i = 10{,}047{,}446$ assuming the largest 9 members of this subpopulation are sampled, one has a = 10,047,446 / 13,657,816 = 73.6%. The 'projected' or 'anticipated' RSE, from equation 5a would be approximately

(4.94 / 1,062,643) $[(13{,}657{,}816)(0.264/0.736)]^{0.5}$ or 1.0-percent.

This is only slightly smaller than the 1.1-percent figure from the quasi-cutoff sample that actually existed. Thus the accommodation for collecting data on multiple attributes did not make a big difference here. This depends upon the situation. For some surveys with many attributes (that is, questions on the survey form), a minor attribute may be sampled completely, or nearly completely as data ancillary to collecting data for other attributes. A scatterplot (x vs y) would show a wide range of representation by population member size. This can help to guard against model-failure, but is far from crucial. See Knaub(2010).

Using equation 6, one might assume that $\sigma_{e_0}^{*} \approx 4.94$ is a good estimate of the standard error of the random factors of the estimated residuals under model-based classical ratio estimation. We can find an estimated coverage needed to obtain an RSE estimate of 2-percent, if our estimate $\sigma_{e_0}^{*}$ is acceptable, and given other total survey error considerations, adequate accuracy needed for decision-making from these data, and resource availability. Because $\sigma_{e_0}^{*} \approx 4.94$ is an estimate obtained from a previous test data set, it may change substantially over time, from data set to data set, and is subject to variance as well. Knaub(2012b) showed some examples, examining this. As in sensitivity analyses for forecasting, one may try more than one value for $\sigma^*_{e_0}$ to see the impact on the estimated sample sizes.

4.2.1.2 Coverage needed for projected RSE estimate of 2-percent

Here, to obtain estimated coverage needs for an estimated RSE of 2-percent, with $\sigma^*_{e_0} \approx 5$, using equation 6, we only need the estimates T* and T_x.

Here $\sum_{i=1}^{N} x_i = 13{,}657{,}816 = T_x$, subject to nonsampling error, and we found T* = 1,062,643. T* is obviously going to change, but not greatly. From equation 6 we may estimate required coverage to be

$\{1+ [((0.02)(1{,}063{,}000)/5)^2 \ / \ 13{,}658{,}000]\}^{-1}$, or about 43-percent coverage.

(0.43)(13,658,000) = 5,872,940. So here it appears that we need a sample with corresponding x-values adding to approximately 5,873,000 to obtain a resulting estimated total, T*, with an estimated RSE of approximately 2-percent.

The subtotal for the three largest x-values is 5,965,840. Therefore, a strictly cutoff sample of size n = 3 would appear to be adequate for an estimated RSE of 2-percent. One must be cautious of the volatility of such a small sample size. Also, if there is a nonresponse, or a response that does not pass data quality checks, then n = 2, and this would be problematic. In addition to variance and bias concerns, there is always the variance of the variance estimate to consider.

In addition, as noted above, many survey statisticians are cautious of nonrandom sampling. However, all concerns need to be balanced in the context of total survey error. Establishment surveys, especially for electric power surveys, have been studied extensively for a number of years at the US Energy Information Administration. Results from quasi-cutoff sampling have been very good. Part of these studies have included comparing totals for twelve months of estimated quasi-cutoff sample-based results to later obtained census results. Further, test data were examined. Also, see Knaub(2001, 2002, 2010 and 2011a).

4.2.1.3 Coverage needed for projected RSE estimate of 0.5-percent

Suppose that one would like an RSE estimate of 0.5-percent. Using equation 6, we have, approximately, $\{1+ [((0.005)(1{,}063{,}000)/5)^2 \ / \ 13{,}658{,}000]\}^{-1}$, or about 92.4-percent coverage.

This means that we need a sample with the corresponding x-values totaling to approximately 12,620,000. The subtotal for the largest 16 x-values in this test data set was 12,648,084. Thus n = 16 out of N = 25 could do that well.

Note that experience has shown that often, when a coverage of perhaps 90-percent or more is required, too many small observations are then collected on a frequent basis, and nonsampling error becomes a greater problem. The CRE is somewhat robust against nonsampling error, as accounting for greater heteroscedasticity would otherwise be advisable (Knaub(2011c)). Trying to collect too many small observations also detracts from data quality efforts when obtaining data for the largest respondents. In this example, to obtain a 92-percent coverage, we need the largest 16 members of the population, leaving only N-n = 9 members to be observed during a less frequent census. (Note that although the same data element for a previous census often makes the best regressor for a current sample, that is not always true.)

4.2.2 For test Case B (municipals)

4.2.2.1 Results for Same Sample Size as for Data Collected

If the cases with the largest n = 11 x-values had been the respondents, then results would have changed as follows:

Using $\sigma^*_{e_0} \approx 10.16$, and the new $\sum_{i=1}^{n} x_i = 20{,}143{,}403$ assuming the largest 11 members of this subpopulation are sampled, one has a = 20,143,403 / 28,712,373 = 70.2%. The 'projected' or 'anticipated' RSE, from equation 5a would be approximately

(10.16 / 2,073,375) $[(28{,}712{,}373)(0.298/0.702)]^{0.5}$ or 1.7-percent.

Once again, for these data, the reduction in the RSE estimate is inconsequential.

4.2.2.2 Coverage needed for projected RSE estimate of 1-percent

Here, to obtain estimated coverage needs for an estimated RSE of 1-percent, with $\sigma^*_{e_0} \approx$ 10, using equation 6, we only need the estimates T* and T_x.

Here $\sum_{i=1}^{N} x_i$ = 28,712,373 = T_x, subject to nonsampling error, and we found T* = 2,073,375. From equation 6 we may estimate required coverage to be

$\{1+ [((0.01)(2{,}073{,}000)/10)^2 \ / \ 28{,}712{,}000]\}^{-1}$, or about 87.0-percent coverage.

(0.87)(28,712,000) = 24,979,440. So here it appears that we need a sample with corresponding x-values adding to approximately 24,980,000 to obtain a resulting estimated total, T*, with an estimated RSE of approximately 1-percent.

The subtotal for the 26 largest x-values is 24,944,210. For the largest 27, it is 25,148,192. Therefore, a strictly cutoff sample of size n = 26 would appear to be adequate for an estimated RSE of close to 1-percent.

However, when n = 26 for a category with N = 61, sample sizes are becoming quite large. When many such categories are considered, and the overall burden for data collection for both the respondents and the agency collecting the data become prohibitive, experience has shown that total survey error is problematic. It has even been the case that dropping the smallest observations in a collected sample, using quasi-cutoff sampling and model-based estimation, has resulted in a lower standard error. Consider if those data had not been collected, and more quality control could have been applied to a smaller sample!

Thus, it would be nice to obtain an estimated RSE of 1-percent or less, but this data set indicates that this may not be practical to do this for this category. RSEs are not designed to measure nonsampling error, but as long as a sample, not a census, is considered, nonsampling error does impact RSE estimates. See Knaub(2007c). Further, with the current budget reductions, sample size reductions are also being sought.

Note that the observed data determine what $\sigma^*_{e_0}$ will be. It is then applied to the part of the population that is out-of-sample, but the estimated values are not the cause of the variance, contrary to what experience has shown to be a popular misconception.

4.3 Balanced Sampling

The population mean x value in Case A is 546,313. If a sample size of 9 is used again, but in a balanced sample (see Brewer(1999)), then the volume coverage can only be *a* = (9)(546,313) / 13,657,816 ≈ 36-percent. (Or note that 9/25 = 0.36.) From equation 5a this gives us an estimated RSE of 2.3-percent, as opposed to the quasi-cutoff result obtained, 1.1-percent. From Knaub(2010), the author argues that bias in a cutoff sample is going to normally be quite limited, this is shown empirically in Knaub(2001), along with variance considerations, and further, Karmel and Jain(1987) had quite satisfactory results as well. (See Karmel and Jain(1987), sections 2.3 and 3.) However, as noted, many statisticians do shy away from cutoff sampling. But here we see that a balanced sample, which would be somewhat comparable to a random sample, gives us a larger standard error, more than double what it was for the quasi-cutoff sample used, which would be hard to justify when estimating the grand total for such a data set. If interest is in the smaller members of the population, specifically, then that is another matter.

Now consider Case B. The population mean x value is 470,695. If a sample of size 11 is used again, but in a balanced sample, then the volume coverage would be (11)(470,695) / 28,712,373 ≈ 18-percent. (This could also be found by dividing 11 by 61, that is, n/N.) From equation 5a this gives us an estimated RSE of 5.6-percent, as opposed to the quasi-cutoff result obtained, 1.8-percent. Thus the estimated RSE is tripled.

4.4 Conclusion for These Test Data

As in most establishment surveys, data in these test cases are quite skewed. (See the Appendix below.) A cutoff sample here is considerably better than a balanced sample with regard to variance of the estimated total. Bias is generally not a practical problem, as discussed in Knaub(2010). Using equation 6 and a cutoff sample will find the smallest sample estimated to achieve a given RSE level. Recall, however, that for relatively large sample sizes, the smallest observations can still have enough nonsampling error to be quite problematic. Also note that the advantage of a cutoff sample over a balanced sample grows with increased skewness of the size variable.

Cutoff, quasi-cutoff, and balanced sampling were shown here. An expected sample size from random sampling could be estimated also. A design-based random sample size would generally need to be substantially larger, as a model takes advantage of regressor (auxiliary) data already available, which is also true for model-assisted design-based sampling. - See Section 4.5 and the Appendix below.

4.5 Summary Table of Key Results

Statistics for Obtained Quasi-Cutoff Sample and for Projected Cutoff and Balanced Samples

Regarding Estimated Residential Electric Sales in Tennessee for December 2012

	Case A: Full-Service Cooperatives				Case B: Full-Service Municipals			
	n	N	a	RSE	n	N	a	RSE
Quasi-Cutoff Sample	9	25	70.1	1.1	11	61	68.8	1.8
Cutoff Sample*	9	25	73.6	1.0	11	61	70.2	1.7
**Cutoff Sample	9	25	73.6	1.0	26	61	87.0	1.0
Cutoff Sample	16	25	92.4	0.5	***	-	-	-
Cutoff Sample	3	25	43.0	2.0	-	-	-	****
Balanced Sample	9	25	36.0	2.3	11	61	18.0	5.6

Note: Volume coverages, a, and RSE estimates above are expressed as percent-values.
*Using the same sample size, n, as in the obtained test data.
**Repeated for Case A for comparison to Case B.
***Prohibitively large. ****Close to that already obtained.

5. Conclusions

Equation 6 can be an effective mechanism for estimating volume coverage needs, when using the classical ratio estimator, and model-based estimation. To arrive at a sample size, a cutoff sample provides the smallest size, n, that will provide the needed coverage. This has to be considered in the context of total survey error (TSE), however, where bias and nonsampling error are concerns. A cutoff or quasi-cutoff sample, with a model-based classical ratio estimator, often seems the best methodology. See Knaub(2007b) and Knaub(2010). Nonsampling error may be reduced, resources are less strained, and bias may not be a substantial concern. See Knaub(2001) as further consideration.

To generalize these results to a complex survey with various strata and attributes of interest might best be handled with simulation or operations research techniques, but closed form solutions might also be useful. Further complications, such as the use of various models and regressors across a population, as in Douglas(2012), might substantially complicate a closed form solution, but a few trial-and-error adjustments might be all that is really needed.

Note that the one-regressor, through the origin, model-based CRE for relative standard error, equation 5a, can be written as $RSE^* = \left(\sigma^*_{e_0}/T^*\right)f(a, T_x)$, where a and T_x are functions of (1) the sum over n and (2) the sum over N. Analogously, equation (4.5) for simple random sampling, in Cochran(1977) may be written as $RSE \approx \left(S/\hat{T}\right)f(n, N)$.

A required sample size to achieve the estimated coverage needed from equation 6 depends upon sampling method. See Section 4.4 above.

Acknowledgements

Thanks to Joel Douglas, Samson Adeshiyan, Carrie Hughes-Cromwick, Jason Worrall, and others for helpful conversations; and to Carrie Hughes-Cromwick for programming to verify some example calculations. Thanks to those who have participated in testing and implementation of the quasi-cutoff (sampling), model-based (estimation) procedures referenced below, over more than two decades. The author is grateful to Ken Brewer, Nancy Kirkendall, and others, for past inspiration.

References

Brewer, K.R.W.(1963), “Ratio Estimation in Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process,” *Australian Journal of Statistics*, 5, pp. 93-105.

Brewer, K.R.W. (1999), “Design-based or Prediction-based Inference? Stratified Random vs Stratified Balanced Sampling. *Int. Statist. Rev*., 67(1), 35-47

Brewer, K.R.W. (2002). *Combined Survey Sampling Inference: Weighing Basu’s Elephants*, Arnold, London.

Cochran, W.G.(1977), *Sampling Techniques*, 3rd ed., John Wiley & Sons.

Douglas, J.R.(2007), “Model-Based Sampling Methodology for the new EIA-923,” http://www.eia.gov/pressroom/presentations/asa/asa_meeting_2007/fall/files/modeleia923.ppt Presented to the American Statistical Association Committee on Energy Statistics, October 18, 2007.

Douglas, J.R.(2012), “Efficiently Utilizing Available Regressor Data Through a Multi-Tiered Survey Estimation Strategy,” Energy Information Administration Intranet, *Papers and Reports*. (To be submitted to *InterStat*, http://interstat.statjournals.net/.)

Holmberg, A., and Swensson, B. (2001), “On Pareto πps Sampling: Reflections on Unequal Probability Sampling Strategies,” *Theory of Stochastic Processes*, Vol. 7, pp. 142-155.

Holmberg, A.(2003) *Essays on Model Assisted Survey Planning*, Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences (40 pages)
http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-3417
http://uu.diva-portal.org/smash/record.jsf?searchId=1&pid=diva2:162729

Karmel, T.S., and Jain, M. (1987), "Comparison of Purposive and Random Sampling Schemes for Estimating Capital Expenditure," *Journal of the American Statistical Association*, Vol.82, pages 52-57.

Kirkendall, N.J.(1992), “When Is Model-Based Sampling Appropriate for EIA Surveys?” *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 637-642.
http://www.amstat.org/sections/srms/proceedings/papers/1992_107.pdf

Knaub, J.R., Jr. (1996), “Weighted Multiple Regression Estimation for Survey Model Sampling,” *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 596-599.
http://www.amstat.org/sections/srms/proceedings/papers/1996_101.pdf.

Knaub, J.R., Jr. (1999), “Using Prediction-Oriented Software for Survey Estimation,” *InterStat*,
http://interstat.statjournals.net/YEAR/1999/abstracts/9908001.php?Name=908001,
August 1999. (Note another version in ASA Survey Research Methods Section proceedings, 1999.)

Knaub, J.R., Jr. (2001), "Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias," *InterStat*, June 2001, http://interstat.statjournals.net/YEAR/2001/abstracts/0106001.php?Name=106001. (Note another version in ASA Survey Research Methods Section proceedings, 2001.)

Knaub, J.R., Jr. (2002), "Practical Methods for Electric Power Survey Data," detailed version, *InterStat*, July 2002, http://interstat.statjournals.net/YEAR/2002/abstracts/0207001.php?Name=207001. (Note shorter version in ASA Section on Government Statistics proceedings, 2002, JSM CD.)

Knaub, J.R., Jr. (2003), "Applied Multiple Regression for Surveys with Regressors of Changing Relevance: Fuel Switching by Electric Power Producers," *InterStat*, May 2003, http://interstat.statjournals.net/YEAR/2003/abstracts/0305002.php?Name=305002

Knaub, J.R., Jr. (2004), "Modeling Superpopulation Variance**:** Its Relationship to Total Survey Error," *InterStat*, August 2004, http://interstat.statjournals.net/YEAR/2004/abstracts/0408001.php?Name=408001 (Note another version in ASA Survey Research Methods Section proceedings, 2004.)

Knaub, J.R., Jr.(2005), "Classical Ratio Estimator," *InterStat***,** October 2005, http://interstat.statjournals.net/YEAR/2005/abstracts/0510004.php?Name=510004

Knaub, J.R., Jr. (2007a), "Heteroscedasticity and Homoscedasticity" in N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. (pp. 431-432). Thousand Oaks, CA: SAGE Publications, Inc.

Knaub, J.R., Jr. (2007b), "Cutoff Sampling and Inference," InterStat, April 2007, http://interstat.statjournals.net/YEAR/2007/abstracts/0704006.php?Name=704006

Knaub, J.R., Jr. (2007c), "Model and Survey Performance Measurement by the RSE and RSESP," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 2730-2736. http://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000197.pdf

Knaub, J.R., Jr. (2008), "Cutoff Sampling," in P. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods.* (pp. 176-177). Thousand Oaks, CA: SAGE Publications, Inc.

Knaub, J.R., Jr. (2009), "Properties of Weighted Least Squares Regression for Cutoff Sampling in Establishment Surveys," *InterStat*, December 2009, http://interstat.statjournals.net/YEAR/2009/abstracts/0912003.php?Name=912003.

Knaub, J.R., Jr.(2010), "On Model-Failure When Estimating from Cutoff Samples," InterStat, July 2010, http://interstat.statjournals.net/YEAR/2010/abstracts/1007005.php?Name=007005.

Knaub J.R., Jr.(2011a). "Cutoff Sampling and Total Survey Error," *Journal of Official Statistics*, Letter to the Editor, 27(1), 135-138, http://www.jos.nu/Articles/abstract.asp?article=271135. (click on "Full Text")

Knaub, J.R., Jr.(2011b), "Some Proposed Optional Estimators for Totals and their Relative Standard Errors for a set of Weekly Cutoff Sample Establishment Surveys," *InterStat*, July 2011, http://interstat.statjournals.net/YEAR/2011/abstracts/1107004.php?Name=107004.

Knaub, J.R., Jr. (2011c), “Ken Brewer and the Coefficient of Heteroscedasticity as Used in Sample Survey Inference,” *Pakistan Journal of Statistics,* Vol. 27(4), 2011, 397-406, invited article for special edition in honor of Ken Brewer’s 80th birthday, found at http://www.pakjs.com/journals//27(4)/27(4)6.pdf.

Knaub, J.R., Jr. (2012a), “Use of Ratios for Estimation of Official Statistics at a Statistical Agency,” *InterStat*, May 2012, http://interstat.statjournals.net/YEAR/2012/abstracts/1205002.php?Name=205002

Knaub, J.R., Jr. (2012b), “Projected Variance for the Model-Based Classical Ratio Estimator,” *InterStat*, September 2012, http://interstat.statjournals.net/YEAR/2012/abstracts/1209001.php?Name=209001

Knaub, J.R., Jr., (2013), "Projected Variance for the Model-Based Classical Ratio Estimator II: Sample Size Requirements," *InterStat*, March 2013, http://interstat.statjournals.net/YEAR/2013/abstracts/1303001.php?Name=303001

Lee, C.R.(2013), “Use of replicate calibration samples in analytical chemistry: uncertainties due to lack of knowledge of heteroscedasticity,” found at http://www.analyt.chrblee.net/calibration/calibscedastpost2.pdf

Lohr, S.L. (2010). *Sampling: Design and Analysis*, 2nd ed., Brooks/Cole, Boston.

Maddala, G.S. (2001), *Introduction to Econometrics*, 3rd ed., Wiley.

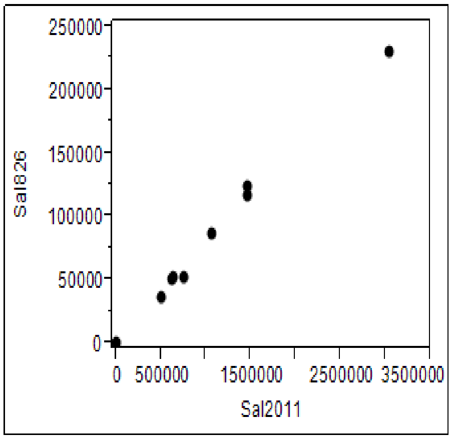
Royall, R.M.(1970), “On Finite Population Sampling Theory Under Certain Linear Regression Models,” *Biometrika*, Vol 57, pp. 377-387.

Royall, R.M., and Cumberland, W.G. (1978), “Variance Estimation in Finite Population Sampling,” *Journal of the American Statistical Association* , Vol. 73, No. 362, pp. 351-358

Appendix: Test Data for Case A: Full-Service Cooperatives

Residential Electric Sales in Tennessee for December 2012 (y), with corresponding regressor data (x) from a census for 2011. A “.” represents missing/out-of-sample y.

Data points on scatterplot		Out-of-sample		Out-of-sample	
x	y	x	y	x	y
3,036,974	229,517	469,220	.	230,963	.
1,470,748	116,450	454,302	.	164,794	.
1,458,118	123,116	434,680	.	162,344	.
1,068,206	85,735	410,281	.	149,989	.
763,099	52,201	383,666	.	34,588	.
643,674	52,371	354,792	.	8,141	.
629,495	50,165	301,780	.	1,006	.
507,912	35,669	261,137	.		
1,457	121	256,450	.		


Sal826
Sal2011