SOME THEORETICAL AND APPLIED INVESTIGATIONS OF MODEL AND UNEQUAL PROBABILITY SAMPLING FOR ELECTRIC POWER GENERATION AND COST

James R. Knaub, Jr., Energy Information Administration
U.S. Department of Energy, EI-541, Forrestal Building, Washington, DC 20585

KEYWORDS: **Auxiliary variate, Ratio estimation, Model-unbiasedness, Probability proportional to size.**

ABSTRACT. The Energy Information Administration (EIA) collects electric power generation and cost data from power plants in the United States. The purpose of this paper is to discuss results of applying model sampling and unequal probability sampling, and to compare these results to each other and to full census historical results where available. This may be used to help determine whether model sampling, or any sampling, is appropriate for some EIA applications. If some of the (smaller) plants are not included in future surveys, this removes some respondent burden, and reduces the number of records that need to be edited, thus possibly improving the quality of the editing and handling for those records that remain. Periodically, censuses may be conducted so that the continuing appropriateness of such methodologies, should any be found, may be examined. This paper represents work-in-progress which may be continued for some time. Also, additional analyses may be conducted.

Introduction. Net generation, fuel stocks, and consumption data are currently collected from every electric utility plant in the United States on a monthly basis. (Such information from nonutility plants is the subject of a separate annual survey which currently does not collect data from the smallest of such plants.) Since census data are available for utility plants, it is possible to compare the results of different sampling estimators to what has resulted from the full census by selecting a subset of that census as the sample. Thus, assuming nonsampling error does not interfere, the difference between sample estimated values and census estimated values will be considered to be due to the mean square error (i.e., bias and variance), and the estimated coefficient of variation (CV) can also be studied for each estimator to see if it behaves as it should. Model sampling is used as that is the most convenient to apply and can remove reporting burden for smaller plants. (Perhaps larger plants will be rotated.)

Due to lack of data, it is usually not possible to investigate the appropriateness of a model so thoroughly. However, this is what the author has done using hydroelectric utility plant generation information. (Data from other plant types have not, at least as yet, been investigated; although data are available.) Historical files are found starting in 1970 for use as auxiliary data. If using prior period generation information, it would be most practical to form an auxiliary data file at the annual level. However, the volatile nature of individual plant data showed that this may not be the best approach.

Net generation may be negative if, for example, a plant is out of service for repairs. A great deal of electricity could be needed for such repairs. Suppose that a plant has a large nameplate capacity (the capacity estimate stamped on the generator by the manufacturer), and for, say, six months in a given year it has a negative net generation, large in absolute value, and for six months it has a large positive net generation, with close to zero net generation for the year. Nameplate capacity would logically be a better auxiliary variate for any given month of that year of interest than prior year generation data might be since a year with near-zero net generation could be chosen for the auxiliary information.

This investigation only considered a recent month's data for 11 States, used nameplate capacity as the auxiliary variate, and covered only hydroelectric plants, at least at this point. Results may vary for years of drought or flood, although such effects may be proportioned well enough over various plants that no substantial net effect may be seen.

The results of using three linear regression models described by Royall (1970) were compared in this study. Total net generation, by State, and CV estimators were compared using census data to determine relative merits (assuming nonsampling error does not bias the results). A fourth model was also studied. The scope of this study was limited, at least at this time, to cases where only the smallest plants were not sampled. This could correspond to a case where only a few utilities are not included due to hardship allowances.

Generation expense estimation is a second part of this paper. Generation expense is a ratio of cost to kilowatthours of electricity generated.[1] The current estimates of net generation, as mentioned earlier, are derived from a census. Thus we must only sample to estimate costs. One of the models in the first part of this paper is used and compared with results from an unequal probability sample design which corresponds partially to another one of the linear regression models used in the first of these two studies. Census results for various cost components are not all available. Plants investigated were major, privately owned coal-fired and nuclear-powered plants for the years 1986, 1987 and 1988. Results are being presented in the current EIA publication, *Electric Plant Cost and Power Production Expenses 1988*.

Model Sampling to Estimate Net Generation For Hydroelectric Plants

Background. As at least a first step in investigating the possible use of model sampling for the collection of net generation information, three models described by Royall (1970) were used to study hydroelectric plant data. These three linear regression models consist of a more general model, and 2 others which adjust for heteroscedasticity (see Maddala (1977), pages 93-94, 259-261).

The models are (Royall (1970) page 378) such that Y_i has mean βx_i and variance $\sigma^2 v(x_i)$, where for Method 1, $v(x_i) = 1$, for Method 2, $v(x_i) = x_i$, and for Method 3, $v(x_i) = x_i^2$. Method 1 corresponds basically to ordinary regression estimation. Method 2 corresponds to ratio estimation and can be found succinctly described, among other places, in Cochran (1977), pages 158-160. Method 3 corresponds somewhat to unequal probability sampling where sampling is in proportion to some measure of size. However, unequal probability sampling is unbiased due to the nature of the design, whereas the purposive selection for this corresponding model is sometimes substantially biased. Cochran (1977), page 160, mentions that a good use of this method is found in Jessen, et.al. (1947). However, Jessen does not separate observed and unobserved y values as Royall does. This can make a substantial difference in an establishment survey due to the predominance of a relatively few respondents.

Finally, a fourth method was applied. It is described in Cochran (1977), pages 199-200. It is even more general than Method 1 as, unlike the three methods taken from Royall (1970), it is not required that the regression pass through the origin. With this model, lowest variance occurs when $\bar{x} = \bar{X}$ (i.e., the sample is "balanced"), whereas Methods 1-3 have lowest variance when the n observations with the largest x values are chosen. Method 4 does not separate observed and unobserved y values.

Note that a promising model may also lead to an imputation procedure. Also, another consequence of this study could be a study of the usefulness of sampling for nonutility generation.

Equations

See Royall (1970), page 382, concerning variance estimation for Methods 1-3. Note that $\hat{Y}$ is an estimated total.

Method 1: *If* $v(x_i) = 1$, then $\hat{\beta} = b_1 = (\sum^n x_i y_i) / (\sum^n x_i^2)$, $\hat{Y}_{b_1} = b_1 X_N + Y_s$, where X_N is the total of the auxiliary variate values for the unobserved plants, and Y_s is the total of the observed y values ($Y_s = \sum_s y_i$), and $\hat{V}(\hat{Y}_{b_1}) =$

$$\hat{\sigma}_1^2(N - n + \frac{(X - n\bar{x})^2}{\sum^n x_i^2}) \text{ where } \hat{\sigma}_1^2 = \sum^n (y_i - b_1 x_i)^2 / (n-1).$$

Method 2: *If* $v(x_i) = x_i$, then $\hat{\beta} = \hat{R} = (\sum^n y_i) / (\sum^n x_i)$,

$\hat{Y}_R = \hat{R}X$, and $\hat{V}(\hat{Y}_R) = \hat{\sigma}_2^2 (X - n\bar{x})X / (n\bar{x})$ where

$\sum^n x_i = n\bar{x}$ and where $\hat{\sigma}_2^2 = \sum^n (1/x_i)(y_i - \hat{R}x_i)^2 / (n-1)$

and $\hat{\sigma}_2^2 = \hat{\lambda}$ in Cochran (1977), page 159.

Method 3: *If* $v(x_i) = x_i^2$, then $\hat{\beta} = b_3 =$

$(1/n)\sum^n (y_i / x_i)$, $\hat{Y}_{b_3} = b_3 X_N + Y_s$, and

$$\hat{V}(\hat{Y}_{b_3}) = \hat{\sigma}_3^2 (\sum^N x_i^2 - \sum^n x_i^2 + (X - n\bar{x})^2 / n)$$

where $\hat{\sigma}_3^2 = \sum^n (1/x_i^2)(y_i - b_3 x_i)^2 / (n-1)$

and $\hat{\sigma}_3^2 = s_o^2$ in Jessen, et.al. (1947), page 370.

Method 4: $\hat{Y}_{lr} = N\left(\bar{y} + b_4(\bar{X} - \bar{x})\right)$

$$\hat{V}(\hat{Y}_{lr}) = N^2 \hat{\sigma}_\varepsilon^2 \left(\frac{N-n}{nN} + \frac{(\bar{X} - \bar{x})^2}{\sum^n (x_i - \bar{x})^2}\right)$$

where $\hat{\sigma}_\varepsilon^2 = \sum^n \frac{\left[(y_i - \bar{y}) - b_4(x_i - \bar{x})\right]^2}{n-2}$, and $\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2$

in Cochran (1977), pages 199-200, and

$$b_4 = \frac{\sum^n y_i(x_i - \bar{x})}{\sum^n (x_i - \bar{x})^2}.$$

$$\hat{V}(\hat{R}) = \frac{\hat{\lambda}(X - n\bar{x})}{Xn\bar{x}} = \frac{\hat{\lambda}}{n\bar{x}}(X - n\bar{x}) / X \quad \text{and}$$

$$(X - n\bar{x}) / X = 1 - (\frac{n}{N})(\frac{\bar{x}}{\bar{X}}) = 1 - f, \quad \text{if} \quad \bar{x} = \bar{X}.$$

However, for $V(\hat{R})$ to be minimized, $\bar{x}$ must be greater than $\bar{X}$.

As another aside, note that from Cochran (1977), page 158, when the ratio estimator is a BLUE:

" 1. The relation between y_i and x_i is a straight line through the origin.
2. The variance of y_i about this line is proportional to x_i. "

Thus, x_i and y_i should be highly positively correlated, and therefore their signs should often be identical.

What happens if $x_{i=k}$ is zero (or nearly zero) and so is $y_{i=k}$? The kth term of $\hat{\lambda}$ should then be zero if the variance of $y_{i=k}$ is to be proportional to $x_{i=k}$. We have

$\lim_{x_k \to 0} (1/x_k)(y_k - \hat{R}x_k)^2 / (n-1)$; and $y_k = \beta x_k + \varepsilon_k$ *so*,

$$\lim_{x_k \to 0} (1/x_k)(\beta x_k + \varepsilon_k - \hat{R}x_k)^2 / (n-1)$$

$$\cong \lim_{x_k \to 0} \left(\frac{(\beta - \hat{R})^2}{n-1}\right) x_k = 0,$$

since ε_k should approach zero much faster than x_k does. However, what if y_k is not near zero when x_k is zero? L'Hospital's rule gives us that the kth term of $\hat{\lambda}$ should be $-2\hat{R}y_k / (n-1)$. This could represent substantial model failure.

Results. The relative performances of Methods 1-4 were studied by several means. Each estimator, $\hat{Y}$, was compared with Y, using the signed rank test. Additional methodology is also described. Some of the data for those comparisons are given in the table below. $\hat{Y}$ and $\hat{CV}$ are the estimated total net generation and estimated CV (percent) values for each method. Y is the total net generation derived from the full census. D is the percent difference between Y and $\hat{Y}$, and z is the approximately standard normal variate derived in the following assuming that a given CV estimate is viable. (Also, Y^* is the estimated total when Y_s is not treated separately from the unobserved y_is. This only affects Methods 1 and 3. In the 11 cases that were studied here, Y_s was a very large portion of Y. It seems reasonable that treating Y_s separately would reduce variance, but the table below also indicates that Method 3 is a poor model for these data unless Y_s is separated. When this is done using Royall (1970), Method 3 performs very well for these data. Apparently the smallest plants related differently to the auxiliary data.)

Now,

$$D = \left(\frac{\hat{Y}-Y}{Y}\right) \times 100\ \%, \text{ and } \hat{CV} = \left(\frac{\hat{\sigma}}{\hat{Y}}\right) \times 100\ \%.$$

$$\left[\left(\frac{Y}{\hat{Y}}\right)\left(\frac{\hat{Y}-Y}{Y}\right) \times 100\ \% \right] \Big/ \left[\left(\frac{\hat{\sigma}}{\hat{Y}}\right) \times 100\ \% \right] =$$

$$\left(\frac{Y}{\hat{Y}}\right) D \,/\, \hat{CV} = \frac{\hat{Y}-Y}{\hat{Y}} \Big/ \frac{\hat{\sigma}}{\hat{Y}} = \frac{\hat{Y}-Y}{\hat{\sigma}} = z_{\hat{Y}}.$$

In the table, $z_{\hat{Y}}$ is simply labeled z.

Thus, for each sample estimate, $\hat{Y}$, and the census value, Y (where a *total* is indicated), a number, z, can be estimated as a standard normal variate. This is done for each methodology for purposes of comparison. Such comparisons make meaningful use of hypothesis tests. Although the hypothesized standard normal distribution in each case is not being compared with another hypothesized distribution function (see Knaub (1987)), results for each estimator are being compared with each other. (See Kolmogorov graphs.)

The Wilcoxon Signed Ranks Test is employed according to Conover (1980), pages 278-292, where the raw data are the values $(Y / \hat{Y})$ and $(Y / Y) = 1$, for each *matched-pair*. This also is done for each method to be compared.

Wilcoxon Signed Ranks Test Results on D
(See Conover (1980), Table A13):

- For *Method 1*: T+ = 13, so fail to reject (two-tailed) at 5 percent, but reject at 10-percent level.
- For *Method 2*: T+ = 10, so fail to reject (two-tailed) at 2 percent, but reject at 5-percent level.
- For *Method 3*: T+ = 28, so fail to reject (two-tailed) at 60 percent, but reject at 80-percent level.
- For *Method 4*: T+ = 47, so fail to reject (two-tailed) at 20 percent, but reject at 40-percent level.

From these results, $\hat{Y}_R$ appears to be the most biased estimator for Y. However, bias is not of any real importance here compared with variance in terms of relative size. The ratio estimate actually appears to be very good, since the absolute difference between Y (census), and $\hat{Y}$ (sample) is least for the ratio estimate, usually by far, for 11 out of 11 data sets. $\hat{Y}_R$, however, can only be said to be best, or second best in this way, 10 of 11 times. $\hat{Y}$ for Method 3 usually did a little better here.

Table 1. Net Generation Data for Five Example Cases

State / Method	$\hat{Y}$	$\hat{CV}$	Y	D	z	Y^*
A						
1	1,807	11.5	1,826	-1.1	-0.09	834
2	1,818	1.1	1,826	-.5	-.45	-
3	1,830	.2	1,826	.2	.81	2,930
4	2,414	10.5	1,826	32.2	2.33	-
B						
1	588.6	8.7	594.0	-.9	-.11	554
2	589.5	2.0	594.0	-.8	-.37	-
3	595.2	1.1	594.0	.2	.18	790
4	612.2	11.1	594.0	3.1	.27	-
C						
1	133.2	7.2	133.3	-.1	-.01	120
2	133.8	1.9	133.3	.4	.20	-
3	134.1	.9	133.3	.6	.66	141
4	142.4	8.6	133.3	6.8	.74	-
D						
1	37.4	50.1	41.2	-9.2	-.20	-66
2	38.1	8.8	41.2	-7.6	-.93	-
3	41.6	1.3	41.2	.8	.63	561
4	89.5	21.2	41.2	117.1	2.54	-
E						
1	6,673	5.5	6,713	-.6	-.11	5,871
2	6,679	.5	6,713	-.5	-.95	-
3	6,693	.1	6,713	-.3	-2.35	8,662
4	7,186	6.4	6,713	7.0	1.03	-

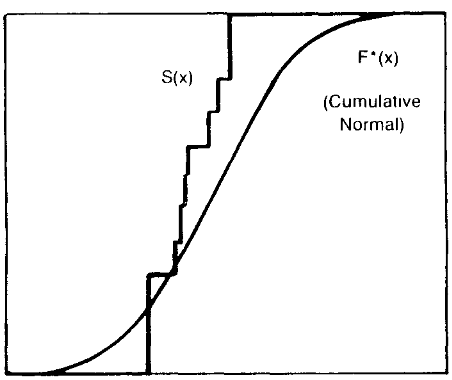
Note: For State A, $\Sigma_s y_i \cong 1797$; for State B, $\Sigma_s y_i \cong 572.9$; for State C, $\Sigma_s y_i \cong 128.1$; for State D, $\Sigma_s y_i \cong 37.9$; and, for State E, $\Sigma_s y_i \cong 6631$.

CV Estimation. In all 11 cases, the absolute value of D for the ratio estimate (i.e., Method 2) was less than the estimated CV. Ideally, the CV estimate will be greater than the absolute value of D with probability approximately 0.68. Here, however, there is a 95-percent confidence that this probability is 72 percent or greater, and this does not consider the fact that the difference was often substantial. (For Method 3, D was less than the estimated *CV* in 8 of 11 cases.)

Kolmogorov-Smirnov (K-S) Testing on z also indicated a substantial bias for the CV estimate for Y_R However, again this is a small problem compared with variance.

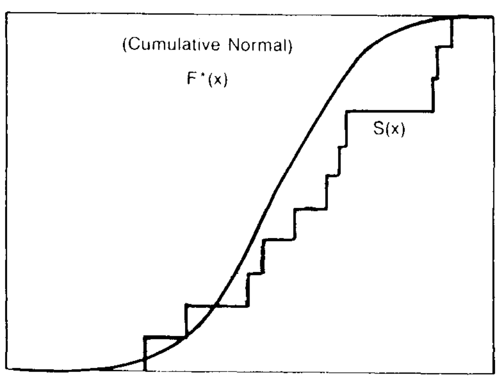
The graphs below are for the K-S test for Methods 2 and 4.

Method 2



T = 0.40 $\Longrightarrow$ Attained two-sided "significance" level approximately 5%
(See Conover (1980), Table A14.)

Method 4



T = 0.26 $\Longrightarrow$ Attained two-sided "significance" level greater than 20%
(See Conover (1980), Table A14.)

It is concluded that although both $\hat{Y}_R$ and $\hat{CV}(\hat{Y}_R)$ appear to be slightly biased, deviations from census based values are generally low -- especially maximum deviations -- for this method. For the cases shown in the table given in this paper, however, Method 3 performs somewhat better, indicating substantial heteroscedasticity dealing with smaller plants.

Note that if we denote $p_i(|z_i|) = 2P[Z < -|z_i|]$, and $\gamma = \Pi p_i$ then for the cases studied, γ(Method 1) > > γ(Method 2) > > γ(Method 3) > > > γ(Method 4). Method 4 sometimes provided poor *CV* estimates.

In 7 of 11 cases, ranking $|D|$ from smallest to largest for these methods, yielded the order 3, 2, 1, 4. There are only approximately 15 chances in 10,000,000 that 1 of 24 possible orderings would occur 7 of 11 times. Thus, for these data, a pattern is shown.

The ratio estimate (Method 2) may be best for general purposes of estimating generation. Note that if the assumptions of this model are strictly correct, estimation may proceed as well with the smaller plants as with the larger ones, as long as more plants are sampled if the smaller ones are used. Suppose the plants in one data set are divided into two groups -- the larger plants and the smaller ones -- and we let $\bar{x}_1$ be $\bar{x}$ for the larger plants, n_1 be the size of that part of the overall sample, and $\bar{x}_2$ and n_2 correspond to the smaller plants. One then wants $n_1 + n_2 = n$ such that $(X - n_1\bar{x}_1) / (X - n_2\bar{x}_2) = (n_1\bar{x}_1) / (n_2\bar{x}_2)$ implies that $n_2\bar{x}_2 = n_1\bar{x}_1$. Establish the two *groups* of plants accordingly. If s_ε is the same in each group, variances will be equal. If Y_R estimates from these two groups are nearly equal, and variance estimates are nearly the same, then this is excellent evidence in support of the model. If a complete census is not available for testing a model, then, in addition to graphical procedures, this method could be adopted to the extent data are available, as a check of the model. Experience on these data, however, shows that it would probably take a rather large data set for enough stability for the above equations to hold approximately true, even if the model works well. Therefore, if these equations do hold approximately true, we have very good evidence that the model is good, and the smaller the sample sizes, the better is such evidence. For the data used, in four cases studied, the Y_R and *CV* values varied greatly, but the absolute values of D did not. This seems to indicate some difference due to plant size, and therefore some model failure, but from the size of the D values in Table 1 for Method 2, the model failure does not appear to be serious. Sample sizes in Table 1 range from less than 50 to more than 200. In some cases, very few plants had most of the net generation. Under such conditions, it is not surprising that the *CV* estimator will be tenuous for the n_1 group described above.

Sampling to Estimate Generation Expense for Major, Privately Owned Coal-Fired and Nuclear-Powered Electric Generating Plants

Background. No census results are available for some of the cost components of generating electricity. These costs are therefore estimated through sampling. The most readily apparent stratification criterion in this new effort was size (net generation, or nameplate capacity), so unequal probability sampling was used and net generation was chosen as the measure of size. Also, model sampling results, using $\hat{Y}_R$ with first net generation and then generator nameplate capacity as the auxiliary variate were studied. For preliminary samples, from which required additional sample sizes were calculated, 4 sets of results were provided: 2 design-based, and the 2 model-based mentioned above. The first design-based method was unbiased, and the second was built upon that. The 2 model-based methods make use of the model-unbiased ratio model shown earlier as *Method 2.*

For the design-based analyses, cost estimates and estimates of their *CV*s were calculated under the design requirement of sampling in proportion to net generation and with replacement. (Negative net generation cases were handled separately.) Next a *without replacement* design-based set of estimates was calculated by not using any duplications of observations, adjusting the inclusion probabilities according to the methodology of Van Beeck and Vermetten shown in Konijn (1973), pages 259-261, and multiplying by the usual finite population correction factor of equal probability sampling, as may be justified by the findings in Cochran (1977), pages 267-270, and Bayless and Rao, as described there by Cochran, and seen to extend to larger sample sizes by comparison to the Rao, Hartley, Cochran Method. There are a number of more exact methods, but they are often quite cumbersome either to calculate or administer, and may not easily accept secondary sampling when the initial sample is found to be inadequate.

The Van Beeck and Vermetten Method resulted in increased relative probabilities of selection for smaller plants over what they were when replacement was allowed. Since the smaller plants may have disproportionately high costs, this could help to lower variance since variance approaches zero as proportionality becomes more exact. If this is an over-adjustment, then at least the result might be like a *Type B* population (found in Cochran (1977), pages 268-269), thus helping to keep variance relatively small. This is apparently because the next best situation to complete proportionality is to have positive correlation between the probability of selection and the mean per element.

Let us consider that n distinct plants were selected, some perhaps multiply, so that the *with replacement* sample size n' is such that $n' \geq n$. If $n' > n$, then systematically chosen subsets of n of the n' observations were used for variance estimation. The mean of such results, using the finite population correction factor (as in Rao-Hartley-Cochran sampling with no remainder term) was used as a *without replacement* type estimate. As stated above, the Van Beeck and Vermetten Method provides inclusion probabilities that may help insure that this correction factor is not too optimistic.

A big advantage of this sampling methodology is the ease with which additional observations may be incorporated. This is convenient if a preliminary sample is taken and the remaining required sample size is then calculated, as was done here. (This procedure is sometimes called *double sampling*.) Also, as was contemplated in this application, this works well if additional observations are wanted in conjunction with poststratification. Also, the model sampling results could be calculated for every corresponding situation.

Note that although only costs are discussed here, costs per kilowatthour have the same *CV* estimates since kilowatthours generated is considered a constant in this application.

In summary, for estimation of generation expense, unequal probability sampling did not remain strictly sampling with probability proportional to size (PPS) after the first observation was drawn, but became proportional to the size of the remaining population on each draw. However, this may have tended to insure (for these data) that variance estimation was not too optimistic, and also means that drawing additional observations was easy. In addition, model sampling was done to see if similar results would follow, bringing the advantages of purposive sample selection to future efforts, and also as a way of comparing the use of net generation and generator nameplate capacity as variates correlated with costs. Results of a poststratification study to separate plants by those with at least one relatively new unit were marginal. Perhaps other criteria may be found. Note also that for the design-based estimates, using the *without replacement* scheme was of limited help since $n' > n$.

Equations and Results

Required sample sizes. Do a preliminary sample to estimate the total size needed. For ratio model sampling, i.e., Method 2 found earlier, let c be a particular *CV* value. Then, $c\hat{Y}_R$ equals the square root of the variance.

Solving for c: $c \cong \dfrac{s_\varepsilon}{\hat{Y}}\left[\dfrac{(X - n\bar{x})X}{n\bar{x}}\right]^{1/2}$

Solving for n: $n \cong s_\varepsilon^2 X^2 / (\bar{x}c^2\hat{Y}^2 + s_\varepsilon^2 X\bar{x})$.

For the *without replacement* design-based sampling found here:

Solving for c: $c \cong \dfrac{s}{\hat{Y}}\left[\dfrac{N - n}{n(N-1)}\right]^{1/2}$

Solving for n: $n \cong \dfrac{N}{\left[(N-1)c^2\hat{Y}^2 / s^2 + 1\right]}$.

If, as happened in the generation expense estimation project, a certainty stratum is used (to account separately for two plants with negative net generation totals for a given year), the last two equations become

$$c \cong \frac{s}{\hat{Y}}\left[\frac{N - n}{(n - n_2)(N - n_2 - 1)}\right]^{1/2}$$

$$n \cong n_2 + \left[\frac{N - n_2}{((N - n_2 - 1)c^2\hat{Y}^2 / s^2) + 1}\right]$$

where n_2 is the number of units (here *plants*) selected with certainty, n is the total sample size, and N is the size of the population.

An example encountered in generation expense estimation follows:

For $N = 63$, $n_2 = 2$ and $n = 14$, the estimate for s was 12,593, and $\hat{Y} = 17,245$. This yields $c \cong 19.1$ percent. If, however, we set $c = 10$ percent, and solve for n, $n \cong 30.7$. Therefore, a total sample size of 29 plus the two certainties was estimated to be needed to bring the CV down to 10 percent. (Thus, there was a lot of variability in the data. Either a better stratification criterion, or set of criteria, must be found, or relatively large sample sizes will continue to be needed. A lot of this variability may actually be due, however, to nonsampling error.)

Let $\hat{V}(PPXWOR)$ be the variance estimate under the *without replacement* unequal probability scheme described above. If we set $\hat{V}(R) = \hat{V}(PPXWOR)$ when sample sizes are equal, and the unequal probability sampling scheme is used, then $n = f(\bar{x}) = ((X/\bar{x})a - Nb)/(a-b)$, where $a = s_\varepsilon^2 X$, and $b = s^2/(N-1)$. (Notice that if $\bar{x} = X/N = \bar{X}$, then $n = N$. This is not likely, however. More likely, $\bar{x} > \bar{X}$.)

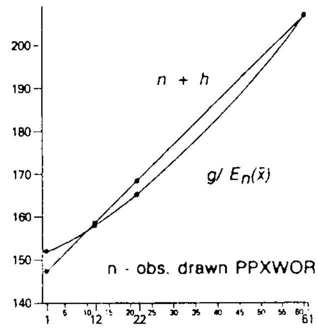
Let $g = Xa/(a-b)$, and $h = Nb/(a-b)$, then $n = f(\bar{x}) = (g/\bar{x}) - h$, and $n > f(\bar{x})$ implies that $n + h > g/\bar{x}$.

Example: $N = 61$, $n = 22$, $X = 453,380$, mean of $s^2 = 120,976,900$, $s_\varepsilon^2 = 6.3006$, $\hat{\bar{x}} = 10,432$, then, $n + h \cong 168$, and $g/\bar{x} \cong 148$, so, $\hat{V}(R) < \hat{V}(PPXWOR)$. However, $E_{n=22}(\bar{x}) \cong 9,329$ (found using Van Beeck-Vermetten inclusion probabilities), thus $g/E_{n=22}(\bar{x}) \cong 165.2$, and $n + h \cong 168.4$.

Since $168.4 > 165.2$, $\hat{V}(\hat{R})$ and $\hat{V}(PPXWOR)$ would be expected to be nearly equal on average, but with $\hat{V}(R) < \hat{V}(PPXWOR)$ in general, when $n = 22$. The graph which follows shows that the point where $n + h$ and $g/E_n(\bar{x})$ cross is near $n = 11$.

Therefore, in the future, the model sampling approach could be at least partially verified by taking an unequal probability sample of (approximately) size 11 as a preliminary sample, and if results are encouraging, only the "largest" plants would be selected to complete the desired sample using a model sampling approach. Little efficiency is lost in this case by waiting until after the preliminary sample selection to decide on the final sampling methodology.

Alternatively, perhaps a census could be used to determine, among other things such as a better stratification possibility, more accurate values for g and h (using generation nameplate capacity as that now appears to be a little better than net generation as a measure of size), and establish an improved estimate of the preliminary sample size for future years. (Variability of s_ε and s estimates for this sample size could be studied also.)



References

- Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed., John Wiley & Sons.
- Conover, W. J. (1980). *Practical Nonparametric Statistics*, John Wiley & Sons.
- Jessen, R. J., et.al. (1947). "On a Population Sample for Greece," *Journal of the American Statistical Association*, pp. 357-384.
- Knaub, J. R., Jr. (1987). "Practical Interpretation of Hypothesis Tests," *The American Statistician*, Vol. 41, p. 246.
- Knaub, J. R., Jr. (1989a). "Ratio Estimation and Approximate Optimum Stratification in Electric Power Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 848-853.
- Konijn, H. S. (1973). *Statistical Theory of Sample Survey Design and Analysis*, North-Holland Pub. Co., and American Elsevier Pub. Co., Inc.
- Maddala, G. S. (1977). *Econometrics*, McGraw-Hill, Inc.
- Royall, R. M. (1970). "On Finite Population Sampling Theory Under Certain Linear Regression Models," *Biometrika*, Vol. 57, pp. 377-387.

Addendum to References

- **I. Hypothesis Testing**
  - Knaub, J. R., Jr. (1989b). "Fellegi-Sunter Record Linkage Theory As Compared to Hypothesis Testing," *Proceedings of the 21st Symposium on the Interface* (of Computing Science and Statistics), pp. 524-527.
- **II. Design-Based Sampling**
  - Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed., John Wiley & Sons.
  - *Contributions to Survey Sampling and Applied Statistics* (1978), Academic Press.
  - *Current Topics in Survey Sampling* (1981), Academic Press.
  - Hansen, M. H., Hurwitz, W. N., Madow, W. G. (1952). *Sample Survey Methods and Theory*, John Wiley & Sons.
  - Jessen, R. J. (1978). *Statistical Survey Techniques*, John Wiley & Sons.
- **III. Model-Based Sampling**
  - Cassel, C-M., Sarndal, C-E., Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*, John Wiley & Sons, Inc.
  - Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed., John Wiley & Sons.
  - *Contributions to Survey Sampling and Applied Statistics* (1978), Academic Press.
  - *Current Topics in Survey Sampling* (1981), Academic Press.

[1]Another EIA survey is used to estimate revenue to the utility per kilowatthour of electricity sold for various (ultimate consumer) sectors of the economy at the State level. Those estimates are made using a double ratio estimate (see Knaub (1989a)) which is design-based rather than model-based. (The author has made plans to investigate the possibility of modeling by comparing design-based results, as they are acquired, to results that are obtained through modeling and the use of data from certainty strata (i.e., generally *large* utilities only).)