SOME SAMPLE DESIGNS EMPLOYING AUXILIARY VARIABLES

Howard L. Jones, 102 Park Lane, Kerrville, TX 78028

Introduction

A number of procedures have been proposed for employing one or more auxiliary variables with known totals for the population as an aid in estimating the unknown total for some variable of interest. Thus, even before they began to use statistical sampling not many years ago, auditors of financial statements commonly used the ratio of error amounts to book amounts in judgment samples of the client's accounts as an estimate of the ratio for the population. And since the sum of the book amounts for the population is always known, this sample ratio could be used to estimate the total dollar error in the accounts.

When auditors got around to using statistical sampling and tried to compute confidence limits by conventional methods, however, they ran into a problem for the reason that the sample often includes very few errors or none at all, and the computed confidence limits were therefore unreliable. See Neter and Loebbecke (1977).

To solve this problem, Anderson and Teitlebaum (1973) propose a procedure which they call dollar-unit sampling. Borrowing an idea from Deming (1960, p. 89), they take each dollar of each book amount (x_j) in the population as a sampling unit. And exploiting the fact that the error amount (y_j) for a particular account is rarely greater than the book amount, they assume 1 to be an upper bound to the prorated error amount (i.e., y_j/x_j) for any dollar unit.

They also employ an unproved theorem in nonparametric estimation to compute an upper confidence limit for the total dollar error in the population. While the authors do not discuss point estimates, it is easy to see that their approach can be used to compute unbiased estimates of the total dollar error in the population.

Unbiased estimates of the variance can also be computed if the dollar units are selected by simple random sampling. If, however, the dollar units are selected by systematic sampling, as the authors propose, such computation would usually not be possible, nor would it be difficult to find exceptions to the unproved theorem just mentioned.

The motivation for the present paper was curiosity as to the efficiency of dollar-unit sampling as compared with that of several other designs employing an auxiliary variable. For study purposes, I chose two simple hypothetical populations, designated Population A and Population B in Table 1 following. Here, z_j is the value of the regression estimate $b_0 + b_1 x_j$ when b_0 and b_1 are computed from data for the entire population. Also, $m_j = 2(z_j + c)$ for Population A, and $m_j = 4(z_j + c)$ for Population B, where c is some positive number, and the factor 2 or 4 is chosen to eliminate fractions.

Capital letter N denotes number of sampling units in the population (3 for Population A and 5 for Population B); while Y, X, Z, and M denote population totals.

Comparison of Mean Square Errors

For each of several sample designs, Table 2 shows the mean square error of estimates of the population total Y when a sample of size 2 is selected from Population A, and when a sample of size 3 is selected from Population B.

The sample design numbered 00 and designated auxiliary-unit sampling is identical to the one called dollar-unit sampling when the auxiliary variable is measured in dollars. This and the other designs shown in Table 2 are discussed briefly in the explanatory notes following.

Of particular interest is the design numbered 5b. Here the selection procedure is **simple random** sampling without replacement, and the estimation procedure is the usual one employing simple linear regression of y on x. However, unlike the procedures numbered 4a to 5a where the regression coefficients for the entire population are assumed to be known, procedure 5b employs coefficients computed from a previous sample of the same size that is statistically independent of the current sample to which these coefficients are applied.

As Table 2 indicates, the mean square error for this last design compares favorably with that for most other designs where the regression coefficients for the population are not given. Moreover, for any combination of regression coefficients that are statistically independent of the current sample, the point estimate $\hat{Y}$ is conditionally unbiased; that is, $E(\hat{Y} | b_0, b_1) = Y$. And a variance estimate that is also conditionally unbiased can easily be computed.

It follows that the statistic $\hat{Y}$ for this design is unconditionally unbiased as an estimate of Y; and the unconditional expectation of the variance estimate is the average of the variances for all possible sample values of the regression coefficients. Also, since the estimates of Y are essentially linear in the variables, perhaps confidence limits computed for this design will be more reliable, and related tests of significance will be more robust, than where the sample design employs nonlinear estimates.

Conclusion

Comparison of the mean square error for Design 5b with that for some other designs shown in Table 2 is a bit unfair, of course, because this particular design requires two samples of size n, while the others require just one sample. In a practical situation, however, where sample surveys of a slowly changing population are carried out periodically, it would seem worth while to investigate the possibilities of this approach where regression coefficients are computed, not from the data for the current survey or from another sample in the current period, but from the data in one or two previous surveys in the not too distant past. In that case, the estimate of Y and the variance estimate will still be unbiased as long as the regression coefficients are statistically independent of the current sample values, even though the population may have changed substantially.

Moreover, when estimates of population totals are desired for each of several variables of interest, we can employ different regression functions of the same or different auxiliary variables, provided values of all such variables are included in the sample surveys used in computing such estimates. And the auxiliary variables used in a particular regression function can include all those that would be used in designing a stratified sample of population values of some particular variable. Thus, for a sample of households, the regression function can incorporate attributes such as urban, rural, western, or irrigated, after quantifying in the usual way by assigning a value of 1 or 0 to a "dummy" variable according to whether an attribute is present or absent.

It is hoped this paper will stimulate further research on these sample designs by statisticians who have access to data for real populations that are sampled periodically.

Explanatory Notes

Design 0. "Circular selection" means selection from the population after arranging the sampling units in a closed circle. See Cochran (1977), Sec. 8.1, with credit to Lahiri.

Design 00. Here X becomes the number of auxiliary units; and the ratio y_i/x_i is the prorated value of the variable of interest for a particular auxiliary unit in the sample. Thus, the sample mean per auxiliary unit for this variable is the mean of such ratios; and expansion of that mean by multiplying by X yields an unbiased estimate of Y.

Design 1. The estimation procedure for Design 1a is simple expansion of the sample mean. Procedure 1b employs a difference estimator. Design 1c employs the usual ratio estimator. Designs 1d and 1e employ unbiased ratio estimators as proposed by Hartley and Ross and by Mickey, respectively, mentioned by Cochran (1977), Sec. 6.15. Procedure 1f employs the usual regression estimator, with coefficients computed from the current sample.

Design 2. For Population A, one sampling unit is selected from Stratum 1, consisting of unit 1 only, and another unit is selected from Stratum 2, consisting of units 2 and 3. For Population B, one unit is selected from each of three strata where Stratum 1 consists of units 1 and 2, Stratum 2 consists of units 3 and 4, and Stratum 3 consists of unit 5 only.

Design 3. The probability of selecting a particular sample combination is made proportional to $\bar{x}$ (or to the sum of the x_i in the sample), so that the usual ratio estimator is unbiased. See Cochran (1977), Sec. 6.15, with credits to Lahiri and Midzuno.

Design 4. The selection and estimation procedures are essentially the same as in Design 3, except that m_i is used in place of x_i. It will be noted that the mean square error (or variance, in this case) appears to approach an asymptotic minimum as $c \to \infty$, in which case the selection procedure obviously approaches simple random sampling.

Design 5. As anticipated, the mean square error for Design 5a is the same as for Design 4e, being equal to $(1 - \rho)^2$ times the mean square error for Design 1a, where ρ is the coefficient of correlation between y and z. Thus, the mean square error for Design 5a is that portion of the variance for Design 1a which is not "explained" by the variable z. For each of Designs 5a and 5b, an unbiased estimate of the mean square error can easily be computed from the squares of sample statistics of the form $y_i - \bar{y} - b_1(z_i - \bar{z})$, provided b_1 is statistically independent of y_i and z_i.

References

Anderson, R., and Teitlebaum, A. D. (1973). "Dollar-unit sampling", C A Magazine (Toronto: Canadian Institute of Chartered Accountants), vol. 102, April, pp. 30-38.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley & Sons, Inc.

Deming, W. E. (1960). Sample Design in Business Research. New York: John Wiley & Sons, Inc.

Neter, J., and Loebbecke, J. K. (1977). "On the behavior of statistical estimators when sampling accounting populations", Journ. Am. Stat. Assn., vol. 72, pp. 501-507.

Table 1. Two Hypothetical Populations

Population A					Population B				
Unit no. j	Variable of interest y_j	Auxiliary variable x_j	Regression estimate z_j	Augmented estimate m_j	Unit no. j	Variable of interest y_j	Auxiliary variable x_j	Regression estimate z_j	Augmented estimate m_j
1	0	1	0.0	2c	1	3	25	1.75	7 + 4c
2	1	2	0.5	1 + 2c	2	0	20	1.00	4 + 4c
3	0	2	0.5	1 + 2c	3	-1	5	-1.25	-5 + 4c
	1 = Y	5 = X	1.0 = Z	2 + 6c = M	4	0	15	.25	1 + 4c
					5	3	35	3.25	13 + 4c
						5 = Y	100 = X	5.00 = Z	20 + 20c = M

Table 2. Comparison of Several Sample Designs

Selection procedure	Estimator of Y	Mean square error Pop. A: n=2	Pop. B: n=3
0. Systematic sampling, circular selection	$N\bar{y}$	.500	3.333
00. Auxiliary-unit sampling			
a. Systematic selection	$X\sum(y_i/x_i)/n$	.250	6.483
b. Simple random selection	$X\sum(y_i/x_i)/n$	.563	18.523
1. Simple random sampling without replacement	a. $N\bar{y}$	.500	11.667
	b. $N(\bar{y} - \bar{x}) + X$	.500	303.333
	c. $X\bar{y}/\bar{x}$	.502	8.520
	d. $\bar{r}X + [n(N - 1)/(n - 1)](\bar{y} - \bar{r}\bar{x})$	.542	11.806
	e. $\bar{r}_{(1)}X + (N - n + 1)(\bar{y} - \bar{r}_{(1)}\bar{x})$	.542	8.308
	f. $Nb_0 + b_1X$ (b_0, b_1 from sample)	.750	4.779
2. Stratified sampling	$\sum N_g\bar{y}_g$	1.000	10.000
3. Ratio sampling	$X\bar{y}/\bar{x}$	.458	7.108
4. Regression sampling	$M(\bar{y} + c)/\bar{m} - Nc$ (b_0, b_1 from population)		
	a. $c = 1/2$	.417	3.049
	b. $c = 1$	.400	2.640
	c. $c = 2$	.389	2.437
	d. $c = 10$	.378	2.308
	e. $c \to \infty$	.375	2.292
5. Simple random sampling without replacement	$N(\bar{y} - \bar{z}) + Z$		
	a. b_0, b_1 from population	.375	2.292
	b. b_0, b_1 from previous sample	.500	3.653