# SOME COMPARISONS BETWEEN LEAST-SQUARES PREDICTION AND UNRESTRICTED RANDOM SAMPLING WHEN THERE ARE TWO CHARACTERISTICS TO BE ESTIMATED

G. W. Lynch, University of Ottawa


## ABSTRACT

For sample sizes 4, 6, 8 and 12, Monte Carlo techniques are used to generate 2,000 random samples (without replacement) from a "real" finite population which has two auxiliary variables, $x_1$ and $x_2$, and two characteristics, $N\bar{Y}_1$ and $N\bar{Y}_2$, to be estimated. The mean square errors (mse) of the population total obtained by these methods are compared to those of the predictive sampling approach. The results indicate that the ratio estimator, under conventional unrestricted random sampling, yield mean square errors which are of the same order of magnitude (for each sample size) when $x_1$ and $x_2$ are used as auxiliary variables; and are decreasing with increasing sample size. Similar results are not obtained under least-square prediction. Additionally, regardless of sample size, unrestricted random sampling is more efficient than the corresponding extreme sample except when information from $x_2$ is used in the estimation of $N\bar{Y}_2$.


## PURPOSIVE SAMPLING

Recently, Royall [1] has presented a methodology, based on least-squares prediction, of sampling from finite populations. The precise sampling scheme is to choose those n units whose x-values are largest (hence an "extreme" or "purposive" sample) and, for this sample, estimate the population total, $Y = N\bar{Y}$, by

$$N\hat{\bar{Y}} = [\sum_s y_j + \hat{\beta} \sum_{\bar{s}} x_j]$$

where the first sum is over the sample units, the second sum is over the units not in the sample,

$$\beta = [\sum_s (x_j y_j / v(x_j))] / [\sum_s (x_j^2 / v(x_j))] ,$$

and $v(x_j)$ is the variance of $x_j$.

When $v(x_j) \alpha x_j$, $\hat{\beta}$ is given by

$$\hat{\beta} = \bar{y}/\bar{x}$$

and $N\hat{\bar{Y}} = N \frac{\bar{y}}{\bar{x}} \bar{X}$.

Thus, in the precise situations for which the ratio estimator is optimal (see Cochran [2]), the classical ratio estimator and the estimator obtained from the predictive sampling approach are identical.

Note also that the "extreme" or "purposive" sample is one of the possible samples under unrestricted random sampling--but it is purposely, not randomly, chosen.

How does purposive sampling compare with unrestricted random sampling? On each of 16 natural populations where there was one characteristic to be estimated, Royall [1] compared the mean square errors obtained under each of the sampling procedures. His results suggest that the predictive sampling scheme generally produced smaller mse's.

It is our contention, however, that multipurpose surveys (rather than unipurpose surveys) are the usual practice. Thus, the natural question to ask is: How will the predictive sampling approach compare with the classical unrestrictive random sampling procedures when there is more than one characteristic to be estimated? To answer this question, we utilized an existing natural population for which there were two quantities to be estimated and computed the mse's under each of the sampling plans for each characteristic.

## THE POPULATION

In a survey conducted in late 1973 (Lynch [3]), we had gathered information from the residents of King and Pierce Counties in the State of Washington. The information on all 350 sample units (N = 350) included: (a) the number of persons in each sample unit ($x_1$), (b) the number of households in each sample unit ($x_2$), (c) the number of females (18 years and older) who had ever had a pap smear ($y_1$), and (d) the number of females, 18 years and older, who had had a recent, 1972 or 1973, pap smear ($y_2$).

## CHARACTERISTICS OF THE POPULATION

TABLE 1: *Population Means, Ranges, Standard Deviations and Coefficients of Variation*

| | Variables | | | |
|---|---|---|---|---|
| | $x_1$ | $x_2$ | $y_1$ | $y_2$ |
| Mean | 9.9 | 3.6 | 3.0 | 2.4 |
| Range | 25 | 8 | 8 | 7 |
| St. Dev. | 5.2 | 1.1 | 1.4 | 1.3 |
| Coeff. of Var. | 0.53 | 0.31 | 0.46 | 0.55 |

$C_{x_1 x_2} = 0.09$, $C_{y_1 y_2} = 0.20$

The population means, standard deviations, ranges and coefficients of variation are presented in Table 1. Here, it is evident that the coefficients of variation range from about 0.1 to approximately 0.55 and that the coefficients of variation of $x_1$, $y_1$ and $y_2$ are approximately equal, while that of $x_2$ is less.

TABLE 2: *Correlation Coefficients*

| Variables | $x_1$ | $x_2$ | $y_1$ | $y_2$ |
|---|---|---|---|---|
| $x_1$ | 1 | 0.58 | 0.65 | 0.54 |
| $x_2$ | 0.58 | 1 | 0.75 | 0.61 |
| $y_1$ | 0.65 | 0.75 | 1 | 0.78 |
| $y_2$ | 0.54 | 0.61 | 0.78 | 1 |

From Table 2 which presents the correlation coefficients, it is evident that all correlations are greater than 0.5. Also, scattergram plots of the data (not shown, in the interest of brevity) revealed that the intercepts were small. Thus, we have the conditions under which the ratio estimator is useful.

## METHODS

Because of issues of bias and variability in small samples, it was decided to cover a range of small sample sizes--that is, n = 4, 6, 8 and 12. Due to limitations on available computer time and financial resources, it was immediately apparent that not all ${}_N C_n$ samples could be generated. Since the computer program, written by Dr. Kronmal [4] and later modified by the author, generated the ${}_N C_n$ samples in a random order, it was decided that a selection of 2,000 random samples for each sample size would provide the desired precision.

For the purposive sampling scheme, we use the n largest units of $x_1$ (Extreme-$x_1$) to compute the mse's for $N\bar{Y}_1$ and $N\bar{Y}_2$. This procedure was repeated for the n largest units of $x_2$.

## RESULTS

The results shown in Table 3 indicate that, when information from either $x_1$ or $x_2$ is used in the estimation of $N\bar{Y}_1$ (see first four rows of Table 1), the purposive sampling plan yields larger mse's at all sample sizes. When $x_1$ was used in the estimation of $N\bar{Y}_2$, the univariate ratio estimator yielded the smallest mse's at all sample sizes; the reverse was true when $x_2$ was employed.

TABLE 3: *Comparison of Mean Square Error Results for the Extreme and the Univariate Ratio Estimators in a Multipurpose Survey*

| Estimator | Mean Square Error, $N\bar{Y}_1$ | | | |
|---|---|---|---|---|
| | n = 4 | n = 6 | n = 8 | n = 12 |
| Ratio -$x_1$ | 60874 | 37811 | 27560 | 16314 |
| Extreme-$x_1$ | 104431 | 106697 | 94080 | 122769 |
| Ratio -$x_2$ | 35664 | 17285 | 12249 | 8096 |
| Extreme-$x_2$ | 31813 | 27360 | 19268 | 10266 |
| | Mean Square Error, $N\bar{Y}_2$ | | | |
| Ratio -$x_1$ | 57087 | 36039 | 27435 | 15681 |
| Extreme-$x_1$ | 134260 | 121250 | 94638 | 101025 |
| Ratio -$x_2$ | 35410 | 23068 | 16706 | 10635 |
| Extreme-$x_2$ | 11519 | 7302 | 2288 | 3382 |

The classical univariate ratio estimator yields mse's which are of the same order of magnitude in the estimation of $N\bar{Y}_1$, and then $N\bar{Y}_2$. Similar results were not always obtained under the purposive sampling scheme (see Extreme-$x_2$).

Thirdly, under unrestricted random sampling, the ratio estimator yields mse's which are decreasing with increasing sample size. This does not appear to be evident under the purposive sampling plan.

## CONCLUSIONS

Of course, one should be cautious about drawing general conclusions from the results of a single population. However, on the basis of estimating two characteristics from this population, the results would seem to suggest that the unrestricted random sampling plan has some desirable properties which are not evident under the purposive sampling scheme.

## ACKNOWLEDGMENTS

I would like to express sincere thanks to Professor D. J. Thompson for his comments and suggestions. Much of this work was done while I was a Graduate Student at the University of Washington and supported by a National (Canada) Student Health Fellowship. Currently, the author is supported by the grant, RD10, of the Government of Ontario.

REFERENCES


1. Royall, R.M. (1970): "On finite population sampling under certain linear regression models." *Biometrics*, 57: 377-387.

2. Cochran, W.G. (1963): *Sampling Techniques. 2nd Edition.* John Wiley and Sons, Inc., New York.

3. Lynch, G.W. (1974): "An evaluation of the accuracy and efficiency of an area sample of King and Pierce Counties." Unpublished M.Sc. Thesis, University of Washington.

4. Kronmal, R. (1975): Personal Communication.