AN ILLUSTRATION OF FINDING OPTIMUM INCREMENT SIZE IN SYSTEMATIC SAMPLING

C. H. Proctor, North Carolina State University

Choice of a sampling unit -- its size and its shape -- is a basic sample design problem [see Cochran, 1972, p. 233 and Hansen, Hurwitz and Madow, 1953, section 6.28], just as is choice of plot size and shape in experimental design. The setting for this choice in the present instance is systematic sampling from a one-dimensional frame. We propose to choose how many adjacent elements to observe at each of the evenly spaced sampled locations. Thus the choice here involves size alone as there is only one shape -- a row.

It is supposed that some data are available to allow us to fit a theoretical one-dimensional stochastic process, a time series, to serve as the superpopulation process. A simple linear cost function will also be taken to be reasonable and for which we have some idea of the cost coefficients. From the correlogram of the process on the elements we show how to transform it to the correlogram for the sampling unit process. By fixing total variable cost we can calculate the systematic sample spacings for different sizes of sampling unit. We show, in passing, how to select a systematic sample when sample size does not exactly divide population size. We then compute process average sampling variances, using the expression for this variance given by Cochran (1946), for various sizes of sampling unit. The recommended size becomes that which makes this variance smallest. Finally, we show how to estimate process variance.

Although our initial work on this problem dealt with sampling the fuel stream of an electric power generating plant (Proctor, 1981) and this explains our use of "increment" in the title, the example we will be using below is the seedling counts data in Cochran's sampling textbook (Cochran, 1977, p. 230). These are the numbers of tree seedlings in each one-foot (about 0.3 m.) of row in a 200 ft. long bed. They are easily accessible data and this may encourage others to compare our approach to theirs.

Frame elements are the one-foot lengths and a sampling unit (SU) will be defined as M adjacent such elements. The case M=1 has N=200; for M=2, N becomes 100; for M=3, we will take population size to be 67 although we set N=66.66 ... in the formulas, and so forth. The cost function is given as:

$$C_T = n(C_1 + C_2 M), \qquad (1)$$

where C_T is total available resource, for example, 30 minutes, C_1 is a per-location cost coefficient and C_2 is a per-element cost, while n is the number of sampled locations, each at N/n SU's apart. Our formulas will work even when k=N/n is not integer. We will simply suppose the results to be close to what more exactly would be found using some gaps of $\lfloor N/n \rfloor$ and others of $\lfloor N/n \rfloor + 1$. The square brackets signify "largest integer in." As a somewhat reasonable cost coefficient we will take C_1 = 1 minute and C_2 = 2 minutes. As a somewhat unreasonable, but interesting, case we will take C_1 = 2.5 and C_2 = 0.5 and also we will take the very unreasonable case of C_1 = 2.8 and C_2 = 0.2.

As stochastic process we use a sum of first order Markov processes which has a mixture-of-exponentials correlogram. Denote the correlation between two elements d units apart as $\rho(d)$. For this process:

$$\rho(d) = \sum_{J=1}^{J} \pi_j \rho_j^d, \qquad (2)$$

with $\Sigma^J \pi_j = 1$. This correlogram shape has been suggested by A. C. Das (1956). From experience we have settled on J = 2 components. To fit this model one first computes the serial correlations r_d for d = 1(1)10(5)100 and then finds values for π_1, and thereby $\pi_2 = 1 - \pi_1$, for ρ_1 and for ρ_2 that produce a fitted $\rho(d)$ that most closely, in generalized least squares, matches the r_d. The serial correlations were found to be: .503, .451, .381, .305, .326, .315, .270, .274, .241, .295, .214, .281, .209, .135, .091, .188, .037, .117, -.002, .023, .025, .200, .146, .135, -.024, -.127, -.085, -.129, -.058, .129, -.017, and .104.

As variance-covariance matrix for the empirical r_d we use a $(\pi, 1-\pi)$ mixture of the two variance-covariance matrices, one based on ρ_1 and the other on ρ_2, for first-order processes as given by Bartlett (1946). This formula is

$$n^{-1}\mathrm{Cov}(r_k, r_{k+d}) = \rho^d(1-2\rho^{2k})(1+\rho^2)(1-\rho^2)^{-1} + d\rho^d - (2k+d)\rho^{2k+d}, \qquad (3)$$

where n is sample size in the pilot survey. The pooled covariances are those in the matrix:

$$\mathcal{V} = \pi \mathcal{V}_1 + (1-\pi)\mathcal{V}_2 \qquad (4)$$

where the entries in $\mathcal{V}_1$ are based on ρ_1 and those in $\mathcal{V}_2$ on ρ_2.

We used starting values based on r_1 as $\hat{\rho}_1^{(1)} = r_1$ and $\hat{\rho}_2^{(1)} = r_1^{10}$ with $\pi^{(1)} = .5$. The (u+1)st values are obtained by settings:

$$\theta^{(u+1)} = \theta^{(u)} + W_u F_u' \Sigma_u^{-1}(\mathbf{r} - \hat{\mathbf{r}}_u) \qquad (5)$$

where the 3 by 1 vectors θ contain the three parameters. F has derivatives of (2) with respect to the parameters; $W = (F'\Sigma^{-1}F)^{-1}$; $\mathbf{r}$ contains the sample serial correlations, and $\hat{\mathbf{r}}_u$ has the fitted values. The subscript u refers to the uth iteration values in $\theta^{(u)}$. For the seedling count data we found $\pi \doteq .48$, $\hat{\rho}_1 = .9297$ and $\hat{\rho}_2 = .0990$. Actually, we fitted to $\lambda_1 = -\log(\rho_1)$ and $\lambda_2 = -\log(\rho_2)$. Also of interest are the alternative quantities $\gamma_1 = 1/\lambda_1$ and $\gamma_2 = 1/\lambda_2$, which may be called "terms" in temporal nomenclature. That is, $\hat{\gamma}_1$ = 14 feet and $\hat{\gamma}_2$ = 0.4 foot (or 5 inches) which distances represents the separation required to bring correlation to $e^{-1} = .37$. The component with $\hat{\gamma}_1$ = 14 feet is the longer term component and $\hat{\gamma}_2$ = 0.4 foot is the shorter term.

The shorter term component may, depending on the applications, reflect, to some extent, measurement error. That is, a measurement error term may have been added to each observation as an uncorrelated sequence of effects or only slightly correlated ones. However, natural processes also exhibit short term correlations. In the present case of counts of numbers of seedlings the short term correlations are more likely natural than arising from the measurement (counting) operation. If we had judged

otherwise then we would need to divide the observed correlations by measurement reliability to correct them for attenuation before fitting.

To transform a correlogram fitted to the element process to one for the sampling unit (SU) process, with M elements per SU, requires a bit of tedious but simple algebra.

It turns out that the resulting correlogram becomes:

$$\rho_M(s) = \Sigma \pi_j^* (\rho^M)^s ,$$

where

$$\pi_j^* = \pi_j (\rho_j^{M+1} - 2\rho_j + \rho_j^{1-M})(1-\rho_j)^{-2} / \left[\sum_{\ell=1}^{J} \pi_\ell (M - M\rho_\ell^2 - 2\rho_\ell + 2\rho_\ell^{M+1})(1-\rho_\ell)^{-2}\right]. \quad (6)$$

Thus the weights are rather messy features to adjust.

Having a correlogram appropriate to the SU process one next applies Cochran's (1946) formula for the variance of a systematic sample from a first order process. We require to apply the modified weights to the two versions of the basic variance expression. That is,

$$\sigma^{-2} V_{SYS} = \Sigma \pi_j^* V_{SYS}(\rho_j) ,$$

where

$$V_{SYS}(\rho) = (k-1)/nk + 2[(n-1)\rho^k - n\rho^{2k} + \rho^{2k+k}]n^{-2} \times (1-\rho^k)^{-2} - 2[(N-1)\rho - N\rho^2 + \rho^{N+1}]N^{-2} \times (1-\rho)^{-2} , \quad (7)$$

and σ^2 is process variance.

The values of V_{SYS} are given in the table for the two cost functions. Under the reasonable cost coefficients of $C_1 = 1$ and $C_2 = 2$ the design with $M = 1$ is best, but if we set $C_1 = 2.5$ and $C_2 = .5$ then $M = 2$ is better than either $M = 1$ or $M = 3$. For the very unreasonable case where $C_1 = 2.8$ and $C_2 = .2$ the optimum SU size is $M = 6$ but the "optimum is very flat." On the other hand the low per element cost permits one to reduce variance from .061 when $M = 1$ to .045 when $M = 6$.

The variance formula is perfectly servicable, as we mentioned earlier, even wehn the cost function calls for k and n values that are not integers. Realizing the design may, however, require some slight struggle. Let us illustrate how to implement the design with $M = 2$ and cost function $30 = n(2.5 + .5(2))$ which seems to imply $n = 8.57$. We round to $n = 9$ and draw a systematic sample of size 9 from the reconstituted frame having $N = 100$ sampling units. To do this we select a random start number in the range 1 to 100 and add the gap, 100/9 = 11.11, successively nine times to the start number, subtracting 100 when needed. This should, barring arithmetic error, return us exactly to the start number. Finally, we round the resulting sequence to produce the selection numbers.

The results thus far on optimal size of sampling unit are unaffected by the size of the process variance σ^2 and thus we have not bothered to estimate the value of σ^2. In applications one might like to know whether a suggested design would attain a sufficiently small variance and thus would require σ^2 to be estimated. If we treat the $N = 200$ values as the finite population Y-values so that $Y_1 = 8$, $Y_2 = 6$, ..., $Y_{200} = 3$ (see Cochran, 1977, p. 230), then $S^2 = \Sigma(Y_i - \bar{Y})^2/(N-1) = 23{,}601/199 = 119.6$. Because of the correlation structure $\underset{\infty}{E}(S^2) = \sigma^2(1-\bar{\rho})$ where $\bar{\rho}$ is the average correlation over all $N(N-1)/2$ pairs of Y-values in the finite population.

In fact:

$$\bar{\rho} = 2 \sum_{u=1}^{N} (N-u)\rho(u)/[N(N-1)]$$

$$= 2\Sigma \pi_j [(N-1)\rho_j - N\rho_j^2 + \rho_j^{N+1}]/[(1-\rho_j)^2 N(N-1)]$$

$$\doteq .058 .$$

The numerical value is obtained by putting our estmates in for π_1, $\pi_2 = 1 - \pi_1$, ρ_1 and ρ_2. Thus we find $\hat{\sigma}^2 = 119.6/(1-.059) = 127.1$.

From the first entry in the table, namely .061, the variance of the one-in-twenty systematic sample would be expected to average, over all finite populations generated by the superpopulation process:

$$\underset{\infty}{E}\, V(\bar{y}_{sy}) = 127.1 \times .061 = 7.75 .$$

This should be compared with the exact variance of the one-in-twenty systematic sample for that particular population, namely 8.19.

References

Bartlett, M. S. (1946). "On the theoretical specification and sampling properties of autocorrelated time series," Jour. Royal Stat. Soc. B: **8**: 27- (Corrigenda, **10**).

Cochran, W. G. (1977). Sampling Techniques, 3rd Edition, New York: Wiley.

Cochran, W. G. (1946). "Relative accuracy of systematic and stratified random samples for a certain class of populations," Annals of Math. Stat. **17**: 164-177.

Das, A. C. (1956). "Two-dimensional systematic sampling and the associated stratified and random sampling," Sankhya **10**: 95-108.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). Sample Survey Methods and Theory, vol. I, Methods and Applications, New York: Wiley.

Table 1. Systematic Sample Variance for Designs with the Same Total Cost[a] but Varying Increment Size and for Differing Cost Coefficients.

Increment or Cluster Size, M	$C_1 = 1$, $C_2 = 2$	$C_1 = 2.5$, $C_2 = .5$	$C_1 = 2.8$, $C_2 = .2$
1	.061	.061	.061
2	.097	.059	.052
3	.143	.062	.048
4	.195	.066	.046
5	.252	.073	.045
6	.310	.080	.045[b]
7	.371	.088	.045
8	.432	.097	.045
9	.493	.106	.046
10	.555	.116	.047

[a] Total cost is $C_T = 30$.

[b] Smallest variance occurs at $M = 6$.