RELATIVE ACCURACY OF SYSTEMATIC AND STRATIFIED RANDOM SAMPLES FOR A CERTAIN CLASS OF POPULATIONS[1]

By W. G. Cochran

Iowa State College

1. Summary. A type of population frequently encountered in extensive samplings is one in which the variance within a group of elements increases steadily as the size of the group increases. This class of populations may be represented by a model in which the elements are serially correlated, the correlation between two elements being a positive and monotone decreasing function of the distance apart of the elements. For populations of this type, the relative efficiencies are compared for a systematic sample of every kth element, a stratified random sample with one element per stratum and a random sample.

The stratified random sample is always at least as accurate on the average as the random sample and its relative efficiency is a monotone increasing function of the size of the sample. No general result is valid for the relative efficiency of the systematic sample. In fact, there are populations in the class in which the systematic sample is more accurate than the stratified sample for one sampling rate, but is less accurate than the random sample for another sampling rate. If, however, the correlogram is in addition concave upwards, the systematic sample is on the average more accurate than the stratified sample for any size of sample.

Some numerical results are given for the cases in which the correlogram is (i) linear (ii) exponential.

2. Introduction. We consider a finite population consisting of the elements $x_1, x_2, \cdots, x_{nk}$, where n and k are integers. A systematic sample is drawn by choosing an element at random from the elements $x_1, \cdots, x_k$, and then selecting every kth consecutive element. That is, if x_i is the element first chosen, the systematic sample comprises the elements $x_i, x_{i+k}, \cdots, x_{i+(n-1)k}$. This type of sample has found considerable use in practice, because it is often easier to select and to administer than a random or stratified random sample and because it has an intuitive appeal through spreading the sample evenly over the population. Much remains to be learned, however, about the accuracy of this systematic sample relative to that of comparable random or restricted random samples. Probably the most relevant comparison is that between the systematic sample and the stratified random sample having one element per stratum. In the latter case, the population is divided into the n strata $\{x_1, \cdots, x_k\}$, $\{x_{k+1}, \cdots, x_{2k}\}, \cdots$, and one element is chosen independently at random from each of the strata. This type of sample is similar in many respects to the systematic

[1] Journal paper No. J-1341 of the Iowa Agricultural Experiment Station, Ames, Iowa. Project 891.

sample. Both divide the population into the same n strata of k elements each, with one element chosen from each stratum. Moreover, neither sample provides the data for an unbiased estimate of the sampling variance of the sample mean, at least in the sense that the estimate is unbiased whatever the form of the population of elements x_i.

The first thorough investigation of the properties of systematic samples was made by W. G. and L. H. Madow [1]. In particular, these authors compared the accuracies of a systematic sample and a stratified random sample of the types described above for several types of finite population. Where the elements in the population lie on the line $x_i = i$, they showed that the stratified random sample, with one element per stratum, is more accurate than the systematic sample. If the population has a periodic distribution, the stratified random sample is superior when k is an integral multiple of the period, but the systematic sample is superior when k is an odd multiple of the half-period. The authors also considered the more complex case where the population contains both a trend function and a periodic function.

The object of this paper is to make similar comparisons for another type of population which appears to be fairly frequently encountered in extensive samplings. The population is one in which the variance among the elements in any group of contiguous elements increases steadily as the size of the group increases. This type of population has long been regarded as applicable in field experimental work, where the variance among plots within a block is found usually to increase with the size of block. Summarizing data from 40 uniformity trials, Fairfield Smith [2] verified this notion and derived an empirical relationship from which the rate of increase may be estimated. The same type of population is also considered in several recent papers on extensive sample surveys. Thus, in a discussion of methods for sampling farm populations, Jessen [3] postulated a law in which the variance among farms within a grid is a monotone increasing function of the size of the grid and used the law for estimating the optimum number of farms which should be included in a sampling-unit. Mahalanobis [4] independently developed the same law as Fairfield Smith in a comprehensive investigation of large-scale sample surveys. Hansen and Hurwitz [5] referred to the increase in variance within a cluster with growing size of cluster as typical of many actual populations. Numerous other references could be given.

3. Specification of the population. Various mathematical models may be constructed to represent the situation in which the variance within any group increases with increasing size of group. For instance, we might consider that the elements x_i are drawn from different populations, the population changing in some regular manner with i. Alternatively, the x_i may be assumed to belong to the same population, but to be serially correlated. For simplicity, we assume further that the serial correlation between x_i and x_{i+u} is some quantity ρ_u which depends only on u. Then if ρ_u is positive and is a monotone decreasing function of u, it may be expected from intuition (and will be proved later) that the variance within the group of elements $x_i, x_{i+1}, \cdots, x_{i+k}$ is a monotone increasing function of k. This model seems appropriate for our purpose, since many writers refer explicitly to positive correlations between the x's as the basis for the phenomenon of increasing variance.

The specification above will be qualified in one respect. To assume that the ρ's are *strictly* monotone for an actual finite population of only moderate size does not seem realistic. While the correlogram may exhibit a definite downward trend, yet individual fluctuations about the trend prevent the correlogram from being strictly monotone. It is more reasonable to regard the finite population as being itself a sample from an infinite population in which the ρ's are monotone. This attitude is, I believe, in accord with that of the authors referred to above, who, as I interpret their writings, regard the variance law as holding in an idealized population. Thus, comparisons between the systematic and stratified random samples will be made not for a single finite population, but for the average of finite populations drawn from an infinite population with monotone decreasing ρ. Results for an individual finite population will differ from the average results because the r's which appear in the population fluctuate about their expectations ρ. As the finite population becomes larger, its results will tend to coincide with the average results.

Accordingly, the elements x_i, $i = 1, 2, \cdots, nk$, are assumed to be drawn from a population in which

$$E(x_i) = \mu,\ E(x_i - \mu)^2 = \sigma^2,\ E(x_i - \mu)(x_{i+u} - \mu) = \rho_u\sigma^2$$

where $\rho_u \geq \rho_v \geq 0$, whenever $u < v$.

4. Some useful preliminary formulas. If $\bar{x}$ is the mean of a specified finite population, the following algebraic identity, frequently useful in the analysis of variance, is easily established.

$$(1) \qquad (kn) \sum_{i=1}^{kn} (x_i - \bar{x})^2 = \sum_{i=1}^{kn} \sum_{j>i} (x_i - x_j)^2.$$

Since there are $(kn)(kn - 1)/2$ possible pairs of values (x_i, x_j), this gives

$$(2) \quad \sum_{i=1}^{kn} (x_i - \bar{x})^2 = \frac{(kn-1)}{2} E(x_i - x_j)^2 = \frac{(kn-1)}{2} E\{(x_i - \mu) - (x_j - \mu)\}^2$$

where E is taken over the finite population. Now expand the quadratic and average over all finite populations. In the $(kn)(kn - 1)/2$ combinations, there are $(kn - 1)$ in which j exceeds i by 1, $(kn - 2)$ in which j exceeds i by 2, and so on. Hence

$$(3) \quad E \sum_{i=1}^{kn} (x_i - \bar{x})^2 = (kn - 1)\sigma^2 \left\{1 - \frac{2}{(kn)\,(kn-1)} \sum_{u=1}^{kn-1} (kn - u)\rho_u\right\}.$$

To obtain the corresponding expectation for the sum of squares within a single stratum of k consecutive elements, we need only replace (kn) by k in (3). Since the result is the same for all n strata, we obtain

$$(4)\qquad E\text{ (S. S. within strata)} = n(k-1)\sigma^2\left\{1 - \frac{2}{k(k-1)}\sum_{u=1}^{k-1}(k-u)\rho_u\right\}.$$

Formula (3) also gives the expected sum of squares within a specified systematic sample if we replace (kn) by n and u by (ku), since there are n elements in the sample and since the correlations between successive elements are $\rho_k, \rho_{2k}, \cdots$ instead of $\rho_1, \rho_2, \cdots$. The result is the same for each of the k systematic samples. Hence

$$(5)\qquad E\text{ (S. S. within systematic samples)} = k(n-1)\sigma^2\left\{1 - \frac{2}{n(n-1)} \cdot \sum_{u=1}^{n-1}(n-u)\rho_{ku}\right\},$$

5. Average variance for a random sample. The symbols σ_r^2, σ_{st}^2, σ_{sy}^2 will be used to denote the average variances of the means of the random, stratified random and systematic samples, respectively, about the mean of the finite population, this average being taken over all finite populations drawn from the infinite population specified in the previous section. Comparisons with the random sample, though not our main purpose, will be included where they are of interest.

For a single finite population, it has been shown by several writers that the variance of the mean of a random sample is

$$(6)\qquad \frac{1}{n}\cdot\frac{(kn-n)}{(kn-1)}\cdot\frac{1}{kn}\sum_{i=1}^{kn}(x_i-\bar{x})^2$$

where $\bar{x}$ is the mean of the finite population.

From (3), we obtain

$$(7)\qquad \sigma_r^2 = \frac{\sigma^2}{n}\left(1-\frac{1}{k}\right)\left\{1 - \frac{2}{(kn)(kn-1)}\sum_{u=1}^{kn-1}(kn-u)\rho_u\right\}.$$

6. Average variance for a stratified random sample. If $\bar{x}_{st}$ is the mean of a typical stratified random sample, the sampling variance of $\bar{x}_{st}$ is by definition

$$(8)\qquad E(\bar{x}_{st}-\bar{x})^2.$$

Consider first the average over a single finite population. Let $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n$ be the means of the n strata, respectively, and let $x_{1j}, x_{2j}, \cdots, x_{nj}$ be the elements selected from the respective strata. Then (8) may be written

$$(9)\qquad \frac{1}{n^2}E\{(x_{1j}-\bar{x}_1)+(x_{2j}-\bar{x}_2)+\cdots+(x_{nj}-\bar{x}_n)\}^2$$

since $$\sum_{i=1}^{n}x_{ij} = n\bar{x}_{st} \text{ and } \sum_{i=1}^{n}\bar{x}_i = n\bar{x}.$$

Take the average over all k^n samples from the finite population. All cross-product terms vanish, since, for example, x_{1j} appears equally often with x_{21}, x_{22}, $\cdots$, x_{2k}. This gives

$$\frac{1}{kn^2}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)^2 \tag{10}$$

for the variance for a single finite population. The sum of squares involved is, of course, simply the sum of squares within strata. Hence, by (4)

$$\sigma_{st}^2 = \frac{\sigma^2}{n}\left(1-\frac{1}{k}\right)\left\{1-\frac{2}{k(k-1)}\sum_{u=1}^{k-1}(k-u)\rho_u\right\}. \tag{11}$$

7. Average variance for the systematic sample. If $\bar{x}_{sy}$ is the mean of a typical sample, the variance for a single finite population is

$$E(\bar{x}_{sy}-\bar{x})^2 = \frac{1}{kn}\{n\Sigma(\bar{x}_{sy}-\bar{x})^2\} \tag{12}$$

where the sum is taken over the k systematic samples. Since the sum of squares among samples is equal to the total sum of squares in the population *minus* the sum of squares within samples, (12) equals

$$\frac{1}{kn}\sum_{i=1}^{kn}(x_i-\bar{x})^2 - \frac{1}{kn}\text{ (S. S. within systematic samples).} \tag{13}$$

To obtain the average over all finite populations we substitute from (3) and (5) for the first and second terms respectively. The result is

$$\begin{aligned}\sigma_{sy}^2 = \frac{(kn-1)}{kn}\sigma^2&\left\{1-\frac{2}{(kn)(kn-1)}\sum_{u=1}^{kn-1}(kn-u)\rho_u\right\}\\ &-\frac{(n-1)}{n}\sigma^2\left\{1-\frac{2}{n(n-1)}\sum_{u=1}^{n-1}(n-u)\rho_{ku}\right\}.\end{aligned} \tag{14}$$

This reduces to

$$\begin{aligned}\sigma_{sy}^2 = \frac{\sigma^2}{n}\left(1-\frac{1}{k}\right)&\left\{1-\frac{2}{kn(k-1)}\sum_{u=1}^{kn-1}(kn-u)\rho_u\right.\\ &\left.+\frac{2k}{n(k-1)}\sum_{u=1}^{n-1}(n-u)\rho_{ku}\right\}.\end{aligned} \tag{15}$$

It should be noted that the formulas and notations above are different from those used by the Madows, who define ρ and σ^2 with reference to a single finite population and discuss the sample variances for a single finite population.

8. Relative accuracies of random and stratified random samples. First, some general comments. From (7), (11) and (15) the relative efficiencies of the three types of sample are seen to depend only on the linear functions of the ρ's which appear in σ_r^2, σ_{st}^2, and σ_{sy}^2. It is easy to verify that in each case the sum of the coefficients of the ρ's is unity. For the random sample, the linear function involves every serial correlation up to lag $(kn - 1)$ with coefficients which decrease linearly as the lag increases and are independent of the size of sample, depending only on $N = (kn)$, the number of elements in the finite population. For the stratified random sample, only serial correlations with lags up to $(k - 1)$ appear, k being the number of elements in the stratum. As presented in (15), the formula for the systematic sample is separated into two linear functions. The first is the same function as appears in the formula for the random sample except that all coefficients are $(kn - 1)/(k - 1)$ times as large. The second, which carries a positive sign, involves correlations where the lag is a multiple of k.

Thus far the formulae require no restrictions on the ρ's. In considering the case where the ρ's are positive and monotone decreasing, the following lemma is helpful.

LEMMA. *If ρ_i, $(i = 1, \cdots, m)$, are positive and monotone decreasing, that is, $\rho_i \geq \rho_{i+1} > 0$ and if $(\alpha_1 + \alpha_2 + \cdots + \alpha_m)$ is zero, the necessary and sufficient conditions that*

$$(16) \qquad L = \alpha_1\rho_1 + \alpha_2\rho_2 + \cdots + \alpha_m\rho_m \geq 0, \qquad \textit{for all admissible sets of } \rho\textit{'s},$$

$$(17) \qquad \textit{are } \alpha_1 + \alpha_2 + \cdots + \alpha_i \geq 0, i = 1, 2, \cdots, (m - 1).$$

For let $\rho_i = \rho_{i+1} + \delta_i$, where by hypothesis $\delta_i \geq 0$. Then if we substitute successively for $\rho_1, \rho_2, \cdots, \rho_{m-1}$ in terms of $\delta_1, \delta_2, \cdots, \delta_{m-1}$, we find

$$(18) \qquad L = \alpha_1\delta_1 + (\alpha_1 + \alpha_2)\delta_2 + (\alpha_1 + \alpha_2 + \alpha_3)\delta_3 + \cdots + (\alpha_1 + \alpha_2 + \cdots + \alpha_{m-1})\delta_{m-1},$$

the final term in ρ_m vanishing because $(\alpha_1 + \cdots + \alpha_m)$ is zero. Since all $\delta_i \geq 0$, the sufficiency of (17) is obvious. Also, if for any i the coefficient of δ_i is negative, we can make L negative by choosing that δ_i as positive and all other δ's as zero. This establishes necessity.

COROLLARY. *If ρ_i are strongly monotone, i.e., $\rho_i > \rho_{i+1}$, and if at least one of the α_i is different from zero, conditions* (17) *are sufficient to establish that L exceeds zero.* For in (18) all the δ's are greater than zero and by (17) none of the δ's has a negative coefficient. Further, the coefficient of at least one of the δ's must exceed zero, otherwise all the α's would be zero. Hence $L > 0$.

We now show that if the ρ_u are monotone decreasing,

$$(19) \qquad L(k) = \frac{2}{k(k - 1)} \sum_{u=1}^{k-1} (k - u)\rho_u$$

is a monotone decreasing function of k. This is the linear function which appears in the variance of the stratified sample.

$$(20) \quad L(k) - L(k + 1) = \frac{2}{k(k - 1)} \sum_{u=1}^{k-1} (k - u)\rho_u - \frac{2}{(k + 1)k} \sum_{u=1}^{k} (k + 1 - u)\rho_u$$

$$(21) \qquad = \frac{2}{k(k^2 - 1)} \sum_{u=1}^{k} (k + 1 - 2u)\rho_u .$$

Since the sums of the coefficients of the ρ_u are unity in $L(k)$ and $L(k + 1)$, the sum is zero in (21). Hence the lemma may be applied. But it is obvious that the sum of the first i coefficients in (21) exceeds zero, since the coefficients are all positive for $u \leq (k + 1)/2$ and all negative for $u > (k + 1)/2$. Hence

$$L(k) - L(k + 1) \geq 0. \tag{22}$$

Further, by the corollary, if the ρ_u are strongly monotone, $L(k)$ is strongly monotone. Since all ρ_u are positive, this result is sufficient to prove that

$$1 - \frac{2}{k(k - 1)} \sum_{u=1}^{k-1} (k - u)\rho_u \leq 1 - \frac{2}{(nk)(nk - 1)} \sum_{u=1}^{nk-1} (nk - u)\rho_u . \tag{23}$$

Consequently, for any size of sample the average variance of the stratified sample cannot exceed that of the random sample. Further, the relative efficiency of the stratified sample to the random sample is monotone increasing with decreasing size of stratum, i.e. with increasing size of sample. There is, of course, nothing unexpected in these results. Equation (22) also establishes the result mentioned in the third section, that with monotone decreasing ρ, the average variance within strata increases steadily as the size of stratum increases. For if $n(k - 1)$ degrees of freedom are assigned to the sum of squares within strata, formula (4) above shows that the average variance within strata is

$$\sigma^2 \left\{1 - \frac{2}{k(k - 1)} \sum_{u=1}^{k-1} (k - u)\rho_u\right\} = \sigma^2\{1 - L(k)\}. \tag{24}$$

9. Comparison of the systematic and random samples. Upon investigation, it is soon evident that no general results can be established about the efficiency of the systematic sample relative to the random samples, unless further restrictions are made on the form of the population. In order to apply the lemma, we find the sums of the first i coefficients of the linear functions of ρ which appear in the variance formulae (7), (11) and (15). By elementary methods these sums are found to be

$$\tag{25} \begin{aligned} \sum\nolimits_r &= \frac{i(2nk - i - 1)}{nk(nk - 1)} \\ \sum\nolimits_{st} &= \frac{i(2k - i - 1)}{k(k - 1)}, \qquad 1 \leq i \leq (k - 1) \\ &\qquad 1 \qquad\qquad\quad , \qquad i \geq k. \\ \sum\nolimits_{sy} &= \frac{i(2nk - i - 1)}{nk(k - 1)} - \frac{rk(2n - r - 1)}{n(k - 1)}, \end{aligned}$$

where r is the integer such that $(r + 1)k > i \geq rk$.

From the lemma, in order to establish $\sigma_{sy}^2 \leq \sigma_{st}^2$, it would be necessary to show that $\Sigma_{sy} \geq \Sigma_{st}$ for any i. Now if i is less than k, so that r is zero, clearly

$$(26) \qquad \sum\nolimits_{sy} > \sum\nolimits_{st} > \sum\nolimits_{r}, \qquad i = 1, 2, \cdots, (k - 1).$$

except when n is 1, in which case all three are equal.

But if i is an integral multiple of k, say rk, we find

$$(27) \qquad \sum\nolimits_{r} = \frac{r}{n}\left[1 + \frac{(n - r)k}{(nk - 1)}\right], \qquad \sum\nolimits_{st} = 1, \qquad \sum\nolimits_{sy} = \frac{r}{n},$$

so that

$$(28) \qquad \sum\nolimits_{st} > \sum\nolimits_{r} > \sum\nolimits_{sy}.$$

Consequently the conditions of the lemma are not satisfied with regard to the systematic sample and no general theorem exists for all populations with monotone decreasing ρ. The result (26) and the corollary show that for any population in this class which has $\rho_u = 0$, $u > (k - 1)$, the systematic sample is more efficient than the stratified random sample. On the other hand, (28) shows that in a population with the first k of the ρ's equal and the rest zero, the systematic sample has a higher variance than a random sample. If these two results are collated for a population with the first j of the ρ's equal and the rest zero, we see that the systematic sample with stratum size j is less accurate than the comparable random sample, while the systematic sample with stratum size $(j + 1)$ is more accurate than the comparable stratified random sample. Although such a population may not occur in practice, the result suggests that the graph of the variance of the mean against the size of sample is unlikely to exhibit the same regularity for the systematic as for the random samples.

10. Populations in which the correlogram is concave upwards. Further investigation shows that the deciding factors in determining the relative accuracies of the systematic and random samples are the second differences of the ρ_u rather than the first differences. The following result will be proved.

THEOREM: *For all infinite populations in which*

$$\rho_i \geq \rho_{i+1} \geq 0, \; i = 1, 2, \cdots, (kn - 1),$$

and

$$\delta_i^2 = \rho_{i-1} + \rho_{i+1} - 2\rho_i \geq 0, \; i = 2, 3, \cdots, (kn - 2),$$

then

$$\sigma_{sy}^2 \leq \sigma_{st}^2 \leq \sigma_r^2$$

for any size of sample. Further, $\sigma_{sy}^2 < \sigma_{st}^2$, *unless* $\delta_i^2 = 0, i = 2, 3, \cdots, (kn - 2)$.

This result can be proved by expressing the linear functions of the ρ_u in terms of second differences and establishing a new lemma applicable to second differences. An alternative approach is simpler and perhaps more instructive.

Since the ρ_u are monotone decreasing, $\sigma_{st}^2 \leq \sigma_r^2$ by the results in section 8. In (13) above, the variance of the mean of a systematic sample for a specified finite population was expressed as

$$(29)\qquad \begin{aligned}&\frac{1}{kn}\sum_{i=1}^{kn}(x_i-\bar{x})^2-\frac{1}{kn}\,(\text{Total S.S. within systematic samples})\\&=\frac{1}{kn}\sum_{i=1}^{kn}(x_i-\bar{x})^2-\frac{1}{n}\,(\text{Average S.S. within a systematic sample}).\end{aligned}$$

A corresponding equation holds for stratified random samples. For if x_{1j}, $x_{2j}, \cdots, x_{nj}$ are the elements of any stratified random sample with mean $\bar{x}_{st}$

$$(30)\qquad \sum_{i=1}^{n}(x_{ij}-\bar{x})^2=\sum_{i=1}^{n}(x_{ij}-\bar{x}_{st})^2+n(\bar{x}_{st}-\bar{x})^2.$$

Now take the average over all k^n samples. This gives

$$(31)\qquad \frac{1}{k}\sum_{i=1}^{kn}(x_i-\bar{x})^2=(\text{Average S.S. within samples})+nE(\bar{x}_{st}-\bar{x})^2.$$

Since the term on the extreme right is n times the variance of the stratified random sample, a result analogous to (29) follows at once.

Consequently, $\sigma_{sy}^2 \leq \sigma_{st}^2$ if the average sum of squares *within* a systematic sample is greater than or equal to that *within* a stratified random sample. Now by (2), with n in place of (kn), each of these averages is equal to

$$(32)\qquad \frac{(n-1)}{2}\,E(x_{ij}-x_{lj})^2$$

where x_{ij}, x_{lj} are the elements in the sample from the ith and the lth strata respectively, the average being taken over all possible pairs of strata.

We consider a fixed pair of strata and let $l - i = u$. For the systematic sample, corresponding elements in the ith and lth strata are always (ku) elements apart. Hence,

$$(33)\qquad E_{sy}\,(x_{ij}-x_{lj})^2=2\sigma^2(1-\rho_{ku}).$$

For the stratified random sample, there are k^2 possible pairs of elements from the two strata. One pair is $(ku - k + 1)$ elements apart, two pairs are $(ku - k + 2)$ elements apart, and so on, the numbers of pairs rising linearly to k and then decreasing linearly to one for the final pair which are $(ku + k - 1)$ elements apart. This gives

$$(34)\qquad E_{st}(x_{ij}-x_{lj})^2=2\sigma^2\left\{1-\frac{1}{k^2}\sum_{i=-(k-1)}^{(k-1)}(k-|i|)\rho_{ku+i}\right\}.$$

Hence, to complete the proof that $\sigma_{sy}^2 \leq \sigma_{st}^2$, it is sufficient to show that

$$(35)\qquad \sum_{i=-(k-1)}^{(k-1)}(k-|i|)\rho_{ku+i}-k^2\rho_{ku}\geq 0$$

for $u = 1, 2, \cdots, (n - 1)$, that is, for any pair of strata. This may be written

$$(36)\qquad \sum_{i=1}^{(k-1)}(k-i)(\rho_{ku+i}+\rho_{ku-i}-2\rho_{ku})\geq 0.$$

But if $\delta^2_{ku} = \rho_{ku-1} + \rho_{ku+1} - 2\rho_{ku}$ is the second central difference it is easy to show that

$$(37) \qquad \rho_{ku+i} + \rho_{ku-i} - 2\rho_{ku} = \sum_{j=-(i-1)}^{(i-1)} (i - |j|)\delta^2_{ku+j} \geq 0,$$

since by hypothesis $\delta^2_j \geq 0$, $j = 2, 3, \cdots, (kn - 2)$. This proves that the variance between the elements of the systematic sample is greater than or equal to that between the elements of the stratified random sample for any fixed pair of strata. The result for the overall average follows. Hence $\sigma^2_{sy} \leq \sigma^2_{st}$. Further, unless $\sigma^2_j = 0$, for all j, clearly $\sigma^2_{sy} < \sigma^2_{st}$, except for samples of one.

The essential point in the proof may be put as follows. The elements in the ith and lth strata are on the average (ku) elements apart for both the systematic and the stratified random sample. When two elements in the latter sample are $(ku + i)$ elements apart, they are less correlated than on the average, since $\rho_{ku+i} \leq \rho_{ku}$, and thus provide more independent information. The variance between the elements exceeds the systematic sample variance by $2\sigma^2(\rho_{ku} - \rho_{ku+i})$. However, such cases are counterbalanced by an equal number of cases in which the elements differ by $(ku - i)$ and the variance is below the systematic sample variance by $2\sigma^2(\rho_{ku-i} - \rho_{ku})$. Because of the concavity of ρ_u, the losses on the average balance or outweigh the gains.

For the population discussed in section 9, in which $\rho_u = \rho$, $u = 1, 2, \cdots, j$, $\rho_u = 0$, $u > j$, we have $\delta^2_j < 0$, $\delta^2_{j+1} > 0$, and $\delta^2_u = 0$ otherwise. This reversal of the sign of the second difference is the explanation for the anomalous behavior of the systematic samples with stratum sizes j and $(j + 1)$.

The theorem above does *not* prove that the relative accuracy of the systematic to the stratified random sample is a monotone function of n, nor even that σ^2_{sy} decreases steadily as n increases. Actually, there are populations in the class for which neither result holds, as will be illustrated in the next section.

So far as practical applications are concerned, the restriction that the ρ_u should be concave upwards may not be severe. For instance, this condition is satisfied when the correlogram is linear, i.e. $\rho_u = (l - u)/l$, this being one type of correlogram which Wold [6] has considered applicable to economic data. Concavity also holds for the function $\rho_u = e^{-\lambda u}$ which Osborne [7] has suggested for forestry and land-use surveys and for the relation $\rho_u = \tanh (u^{-3/5})$ which Fisher and Mackenzie [8] used for expressing the correlation between the weekly rain at two weather stations as a function of their distance apart. In fact, if ρ_u is conceived of as positive and continuous for all u, a concave upwards function suggests itself naturally.

11. Linear correlograms. It may be of interest to present some results obtained when the correlogram is (i) linear, (ii) exponential, since both types have been suggested as possible models for populations occurring in practice.

In the linear case,

$$(38) \qquad \rho_u = (L - u)/L, u \leq L; \quad \rho_u = 0, u > L.$$

If $L \geq (nk - 1)$, the correlogram is a *straight* line throughout the whole range of the finite population. Since all second differences are zero in this case, we may expect $\sigma_{sy}^2 = \sigma_{st}^2 < \sigma_r^2$. If $L < (nk - 1)$, all second differences vanish except δ_L^2, which is positive. Hence we may expect $\sigma_{sy}^2 < \sigma_{st}^2 < \sigma_r^2$.

The results for these cases are found by elementary summations from the basic formulae (7), (11) and (15). Details of the summations will not be presented. For $L \geq (nk - 1)$, we find

$$(39) \quad \sigma_{sy}^2 = \sigma_{st}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\frac{(k + 1)}{3L}; \qquad \sigma_r^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\frac{(nk + 1)}{3L}.$$

The ratio σ_r^2/σ_{sy}^2 is $(nk + 1)/(k + 1)$, which is approximately equal to n, the size of sample, unless the percentage sampled is large. Thus very large gains in efficiency over random sampling are obtained.

If $L < (nk - 1)$, the formulae are less simple. Consider first $k \geq L$; that is, cases where the percentage sampled is less than $100/L$. If $N = nk$,

$$(40) \qquad \sigma_r^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left\{\frac{3N(N - L) + (L^2 - 1)}{3N(N - 1)}\right\}$$

$$(41) \qquad \sigma_{st}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left\{\frac{3k(k - L) + (L^2 - 1)}{3k(k - 1)}\right\}, \qquad k \geq L$$

$$(42) \qquad \sigma_{sy}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left\{\frac{3N(k - L) + (L^2 - 1)}{3N(k - 1)}\right\}, \qquad k \geq L.$$

It is clear on inspection that $\sigma_{sy}^2 < \sigma_{st}^2$; moreover, it is easy to show that the efficiency of systematic relative to stratified random sampling increases steadily as the size of sample increases.

When the size of sample is increased further so that $k \leq L$, formula (40) remains unchanged, while σ_{st}^2 is now given by the same formula as in (39). The formula for σ_{sy}^2 is more complex. If q is the integral part of the quotient when L is divided by k and r is the remainder, so that $L = (qk + r)$, the formula may be written

$$(42') \qquad \sigma_{sy}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right) \cdot \left\{\frac{qk(k^2 - 1) + 3rk(n - q)(k - r) + r(r^2 - 1)}{3NL(k - 1)}\right\}, \qquad k \leq L.$$

It is noteworthy that the last two terms in the numerator inside the curly bracket vanish whenever L is exactly divisible by k. Further, the second term is of order $nk = N$ and, when present, exerts a much greater weight than the first term. Thus σ_{sy}^2 takes a sudden dip whenever L is a multiple of k. In fact, for $L = qk$, (42′) reduces to

$$(43) \qquad \sigma_{sy}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\frac{(k + 1)}{3N}, \qquad L = qk,$$

so that the variance goes to zero if N is sufficiently large. By comparison with formula (39) for σ_{st}^2 we see that when $L = qk$ the relative efficiency of systematic to stratified random sampling is N/L, which increases beyond bound if N is sufficiently large. In intermediate cases, when the remainder r does not vanish, the leading term in the relative efficiency for N large is $(k^2 - 1)/3r(k - r)$. This varies somewhat irregularly, depending on the relation between L and k,

To illustrate, numerical values are given below when $L = 10$ and the finite population is large enough so that terms in $1/n$ are negligible.

The quantities v_{st}, v_{sy} are the corresponding variances apart from a factor σ^2/N. The stratified sample variance decreases steadily with increasing percentage sampled. On the other hand the systematic sample variance goes to zero and the relative efficiency to infinity when k is 2, 5 or 10. Moreover, in the intermediate cases $k = 3, 4, 6, 7, 8, 9$, the variance and the relative efficiency show no consistent relation to the percentage sampled. For samples of less than 10 per cent, including the cases outside the limits of the table, the relative efficiency decreases steadily from 4 at $k = 11$ to 1 when k is large.

TABLE 1

Variances except for a factor σ^2/N and relative efficiency for systematic and stratified random samples for a linear correlogram

k	2	3	4	5	6	7	8	9	10	11	20
% Sampled	50	33	25	20	17	14	12	11	10	9	5
v_{st}	.10	.27	.50	.80	1.17	1.60	2.10	2.67	3.30	4.00	11.65
v_{sy}	0	.20	.40	0	.80	1.20	1.20	.80	0	1.00	10.00
v_{st}/v_{sy}	∞	1.33	1.25	∞	1.46	1.33	1.75	3.33	∞	4.00	1.16

12. Exponential correlograms. For the exponential $\rho_u = e^{-\lambda u}$ the results are much more regular. Each of the linear functions of the ρ's consists of a finite number of terms of an expansion of the form $(1 - x)^{-2}$. If

$$(44) \qquad f(N, \lambda) = \frac{2}{N(N-1)}\left\{\frac{(N-1)e^{\lambda} - N + e^{-(N-1)\lambda}}{(e^{\lambda} - 1)^2}\right\}$$

which is the sum for σ_r^2, we find

$$(45) \qquad \sigma_r^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\{1 - f(N, \lambda)\}$$

$$(46) \qquad \sigma_{st}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\{1 - f(k, \lambda)\}$$

$$(47) \qquad \sigma_{sy}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left\{1 - \frac{(N-1)}{(k-1)}f(N, \lambda) + \frac{k(n-1)}{(k-1)}f(n, k\lambda)\right\}.$$

It may be shown that the variance of the systematic sample decreases steadily and its efficiency relative to stratified sampling increases steadily as the sample becomes larger.

In order to obtain some idea of the magnitude of the gain in efficiency, consider the case where k and n are large. For this case the relative efficiency, which actually is a function of k, n and λ, turns out to depend almost entirely on the single quantity $(k\lambda)$; or, equally, on the correlation $e^{-k\lambda}$ between the items in successive strata in the systematic sample. If $t = (k\lambda)$, we obtain $\sigma_r^2 = \sigma^2/n$,

$$\sigma_{st}^2 = \frac{\sigma^2}{n}\left\{1 - \frac{2}{t} + \frac{2}{t^2} - \frac{2e^{-t}}{t^2}\right\}, \tag{48}$$

$$\sigma_{sy}^2 = \frac{\sigma^2}{n}\left\{1 - \frac{2}{t} + \frac{2}{(e^t - 1)}\right\}. \tag{49}$$

The relative efficiency is given in Table 2 for a selection of values of e^{-t}, the correlation between the items in successive strata.

The relative efficiency has a limiting value 2 when ρ tends to 1 and decreases slowly towards 1 as ρ falls to zero. The gains in efficiency are quite substantial if ρ exceeds 0.1.

TABLE 2

Relative efficiency of systematic and stratified random samples for an exponential correlogram

ρ	.9	.8	.7	.6	.5	.4	.3	.2	.1
$\sigma_{st}^2/\sigma_{sy}^2$	1.96	1.90	1.84	1.78	1.71	1.64	1.55	1.46	1.33

It was pointed out in section 1 that no unbiased estimate of error is available from a single sample for either the systematic or the stratified random sample. This does not mean that no estimate of error can be attempted. However, any estimate must depend on certain assumptions about the form of the population which is being sampled and is likely to be vitiated insofar as these assumptions are false. If, for instance, the correlogram were assumed to be exponential, formula (47), or (49) in the particular case with n, k large, would appear to be the appropriate basis for the estimation of error from a single systematic sample. Consider the simpler case in which (49) is valid. The correlation between successive items in the systematic sample provides an estimate of e^{-t} and hence of t. Also, if terms in $1/n$ are negligible, the mean square within the systematic sample is found to be an unbiased estimate of σ^2. By substitution in (49) a consistent estimate of the variance of a single systematic sample would be secured, provided that the exponential assumption were correct. The gains in efficiency over stratified and random sampling could also be estimated.

REFERENCES

[1] W. G. and L. H. Madow, "On the theory of systematic sampling," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 1-24.

[2] H. Fairfield Smith, "An empirical law governing soil heterogeneity," *Jour. Agr. Sci.*, Vol. 28 (1938), pp. 1–23.

[3] R. J. Jessen, "Statistical investigation of a sample survey for obtaining farm facts," *Iowa Agr. Exp. Sta., Res. Bull.*, No. 304, (1942).

[4] P. C. Mahalanobis, "On large scale sample surveys," *Roy. Soc. Phil. Trans.*, B231 (1944), pp. 329–451.

[5] M. H. Hansen and W. N. Hurwitz, "On the theory of sampling from finite populations," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 333–362.

[6] H. Wold, "A study of the analysis of stationary time series," *Uppsala* (1938).

[7] J. G. Osborne, "Sampling errors of systematic and random surveys of cover-type areas," *Amer. Stat. Assoc. Jour.*, Vol. 37 (1942), pp. 256–270.

[8] R. A. Fisher and W. A. Mackenzie, "The correlation of weekly rainfall," *Quart. Jour. Roy. Met. Soc.*, Vol. 48 (1922), pp. 234–245.