ON THE THEORY OF SYSTEMATIC SAMPLING, III. COMPARISON OF CENTERED AND RANDOM START SYSTEMATIC SAMPLING[1]

By William G. Madow

University of Illinois

1. Summary. The main result obtained is the following: If a population has monotone decreasing correlogram, then centered systematic sampling is more efficient than random start systematic sampling. It is also shown that if a population is monotonic, then centered systematic sampling is more efficient than random start systematic sampling, but here it is easy to cite cases in which stratified random sampling is more efficient than either. Thus, centered systematic sampling is more efficient than random start systematic sampling, in the conditions (namely, concave upwards and decreasing correlogram) in which Cochran [1] proved that random start systematic sampling is more efficient than stratified random sampling.

2. Introduction, Types of Sampling Considered. In this paper, we discuss the theory of centered systematic sampling technique. As is well known, this technique of selecting samples has long been of practical importance. The theory of centered systematic sampling should also be valid for random start systematic sampling with end-corrections (see Yates [5]) since the latter technique in effect reduces random start systematic sampling to centered systematic sampling.

Inasmuch as the approach used in the demonstrations follows that of earlier papers by Cochran [1] and the present author [3], [4], notation and proofs are presented in condensed form.

The elements of the population are $x_1, x_2, \cdots, x_N$ where $N = kn$. The objective is to estimate $\bar{x}$, the arithmetic mean of the population, on the basis of a sample of size n.

The random start systematic sampling estimate, $\bar{x}_{sy}$, is the arithmetic mean of the n elements obtained by selecting one element by an equal probability selection method from $x_1, \cdots, x_k$ and including in the sample every *kth* element thereafter. The arithmetic means of these k possible samples are denoted by $\bar{x}_1, \cdots, \bar{x}_k$, where $\bar{x}_i$ is the mean of the sample whose first element is x_i. The variance of $\bar{x}_{sy}$ is denoted by σ^2_{sy} expressed in terms of the elements of the population.

If k is odd, the centered systematic sampling estimate, $\bar{x}_c$, is $\bar{x}_{(k+1)/2}$ and if k is even we arbitrarily define, $\bar{x}_c = \bar{x}_{k/2}$. (Actually, if k is even, one might either select $k/2$ or $(k + 2)/2$ at random, or one might designate other patterns for selecting the sample elements instead of, as above, designating the elements $x_{k/2}$, $x_{2k/2}, \cdots, x_{nk/2}$. For example, $x_{k/2}, x_{(2k+2)/2}, x_{3k/2}, x_{(4k+2)/2}, \cdots$ would be prefer-

Received 7/12/52, revised 11/3/52.

[1] Research under a contract with the Office of Ordnance Research, U.S. Army.

able in a monotone population. For our present purposes, it is not important to try to determine the best pattern.) The mean square of $\bar{x}_c$ about $\bar{x}$ is denoted by σ_c^2.

In stratified random sampling we consider $x_{1+(j-1)k}$, $x_{2+(j-1)k}$, $\cdots$, x_{jk} to constitute a stratum, $j = 1, \cdots, n$. Hence, there are n strata each consisting of k elements. We suppose that one element is selected from each of the n strata by an equal probability selection method. The sample mean is denoted by $\bar{x}_{st}$ and the variance of $\bar{x}_{st}$ is denoted by σ_{st}^2 expressed in terms of the elements of the population.

We use E to denote the taking of the expected value when the elements of the population are considered to be constants and $\mathcal{E}$ to denote the taking of the expected value when the elements of the population are considered to be random variables.

Inasmuch as we shall be using the word correlogram somewhat loosely in the following, the word is now discussed. If $x_1, \cdots, x_N$ is an ordered sequence of random variables, if ρ_δ is the correlation coefficient of two random variables whose subscripts differ by δ (e.g., $\rho_2 = \sigma_{x_1x_3}/\sigma_{x_1}\sigma_{x_3}$, and if the correlation is to depend only on δ, then the function $f(\delta) = \rho_\delta$, $\delta = 1, \cdots, N-1$, is often called the correlogram of the sequence. It is usually assumed that the random variables have identical mean values and identical variances. However, when we use the word correlogram in the following, it will refer only to the expected value of the product x_ix_h which we will assume to depend only on $\delta = |i - h|$. Thus, if the random variables have identical mean values our statements refer to the usual correlogram but otherwise the condition we state does not assume the identity of the mean values of the random variables.

3. Monotone populations. Hotelling and Solomons [2] proved that for any quantities $z_1, z_2, \cdots, z_g$, the following inequality is valid

$$\frac{g(\text{median} - \text{arithmetic mean})^2}{\sum_{i=1}^{g}(z_i - \bar{z})^2} \leqq 1 \tag{3.1}$$

if all terms are finite and the denominator does not vanish, where, if g is odd the median has the usual definition, and if g is even, and z' and z'' are the two central quantities, then the median can be any quantity such that $z' \leqq \text{median} \leqq z''$. (The details of their proof are given only for odd g but follow at once for even g.)

If a population is monotone, then either $\bar{x}_1 \leqq \bar{x}_2 \leqq \cdots \leqq \bar{x}_k$ or $\bar{x}_1 \geqq \bar{x}_2 \geqq \cdots \geqq \bar{x}_k$. Hence, if k is odd, $\bar{x}_c$ is the median of $\bar{x}_1, \cdots, \bar{x}_k$ and, if k is even, $\bar{x}_c$ is such that $\bar{x}_{k/2} \leqq \bar{x}_c \leqq \bar{x}_{(k+2)/2}$ or $\bar{x}_{k/2} \geqq \bar{x}_c \geqq \bar{x}_{(k+2)/2}$. Then (3.1) becomes

$$\frac{(\bar{x}_c - \bar{x})^2}{\sigma_{sy}^2} \leqq 1. \tag{3.2}$$

Since $(\bar{x}_c - \bar{x})^2$ is the mean square error of $\bar{x}_c$ about $\bar{x}$ when the elements of the population are not random variables, we have proved:

THEOREM 1. *If the population is monotone, then centered systematic sampling is more efficient than random start systematic sampling.*

Of course, if a population is monotone, and if the size of sample is sufficiently large, then stratified random sampling may be more efficient than centered systematic sampling, since the latter estimate may have a bias that does not tend to zero sufficiently rapidly as the size of sample increases.

In practise, however, even if a population is monotone, centered systematic sampling will often be more efficient than stratified random sampling. To see how this will occur let us define the average variance and average covariance terms of σ_c^2.

Let

$$\bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{i+(j-1)k}, \qquad j = 1, 2, \cdots, n.$$

Then $\sigma_c^2 = S + C$, where,

$$S = \frac{1}{n^2}\sum_{j=1}^{n} (x_{a+(j-1)k} - \bar{x}_j)^2,$$

$$C = \frac{1}{n^2}\sum_{\substack{j,m=1 \\ j\neq m}}^{n} (x_{a+(j-1)k} - \bar{x}_j)(x_{a+(m-1)k} - \bar{x}_m),$$

and $a = (k+1)/2$ if k is odd, $a = k/2$ if k is even. We call S the average variance term and C the average covariance term of σ_c^2.

By the result of Hotelling and Solomons

$$(x_{a+(j-1)k} - \bar{x}_j)^2 \leqq \frac{1}{k}\sum_{i=1}^{k} (x_{i+(j-1)k} - \bar{x}_j)^2, \qquad j = 1, \cdots, n.$$

Hence $S \leqq \sigma_{st}^2$. Thus, if $C < \sigma_{st}^2 - S$ then $\sigma_c^2 < \sigma_{st}^2$. In practise, the average covariance term, C, is often small enough for the above condition on C to be satisfied.

4. Populations with monotone decreasing correlograms. (Actually, it is terms such as (4.1) below that will be assumed to be monotone decreasing.)

We will need the following notation in this section. Unless specific limits are stated, the letters i, h will assume all integral values from 1 through k; the letters j, m will assume all integral values from 1 through n; the letter γ will assume all integral values from 1 through $n-1$; the letter δ will assume all integral values from 1 through $k-1$; and the letter ϵ will assume all integral values from 1 through $\frac{1}{2}(k-1)$. (In the proof k is assumed to be odd. The case where k is even introduces further complications and notation without altering the basic results.) We now suppose that the elements of the population are random variables, and let

$$\mathcal{E}x_{\alpha+(j-1)k}x_{\beta+(m-1)k} = \mu_{(m-j)k+\delta}, \tag{4.1}$$

where $\delta = \beta - \alpha$ and $j \leqq m$. Thus, $-(k-1) \leqq \delta \leqq (k-1)$.

THEOREM 2. *Under the conditions stated*

$$\mathcal{E}\sigma_{sy}^2 - \mathcal{E}\sigma_c^2 = \frac{4}{n^2k^2}\sum_{\epsilon}\epsilon\sum_{\gamma=0}^{n-1}(\mu_{\gamma k+\epsilon} - \mu_{(\gamma+1)k-\epsilon}). \tag{4.2}$$

If $\mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_{nk-1}$ *and the inequality holds at least once, then centered systematic sampling is more efficient than random start systematic sampling, while if* $\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_{nk-1}$ *and the inequality holds at least once, then the contrary is true.*

Before proving Theorem 2 let us consider some of its implications. Actually from (4.2) it follows that $\mathcal{E}\sigma_{sy}^2 = \mathcal{E}\sigma_c^2$ if the elements of the population have the same expected product (4.1) no matter how distant they are, that is if $\mu_1 = \mu_2 = \cdots = \mu_{nk-1}$. If we assume all elements of the population have the same expected value then the above statement is made for the serial covariance rather than the expected product. For example, if the elements of the population have the same expected values and are uncorrelated, then $\mathcal{E}\sigma_{sy}^2 = \mathcal{E}\sigma_c^2$.

Furthermore the conditions stated above under which centered systematic sampling is less efficient than random start systematic sampling should almost never be satisfied in practise. In practise, however, irregularities of the correlogram may well lead to the greater efficiency of random start systematic sampling as compared with centered systematic sampling.

PROOF. The demonstration of (4.2) is tedious, but not difficult. Let us begin by obtaining the following two lemmas.

LEMMA 1. *If* $f(i-h)$ *is a function of the difference of the integers* i *and* h, *then*

$$\sum_{i,h} f(i-h) = kf(0) + \sum_{\delta}(k-\delta)[f(\delta) + f(-\delta)]. \tag{4.3}$$

Also,

$$\sum_{i,h} f(|\, i-h\,|) = kf(0) + 2\sum_{\delta}(k-\delta)f(\delta). \tag{4.4}$$

The proof is omitted.

LEMMA 2. *Let* $\delta = |\, i-h\,|$. *Then, if* $i \neq h$

$$\mathcal{E}\bar{x}_i\bar{x}_h = \frac{\mu\delta}{n} + \frac{1}{n^2}\sum_{\gamma}{}'(n-\gamma)[\mu_{\gamma k+\delta} + \mu_{\gamma k-\delta}], \tag{4.5}$$

and

$$\mathcal{E}\bar{x}_i^2 = \frac{\mu_0}{n} + \frac{2}{n^2}\sum_{\gamma}(n-\gamma)\mu_{\gamma k}. \tag{4.6}$$

PROOF. We now denote $x_{i+(j-1)k}$ by x_{ji}. Since $\bar{x}_i = (1/n)\sum_j x_{ji}$, it follows that

$$\mathcal{E}\bar{x}_i\bar{x}_h = \frac{1}{n^2}\sum_j \mathcal{E}x_{ji}x_{jh} + \frac{1}{n^2}\sum_{j<m} \mathcal{E}x_{ji}x_{mh} + \frac{1}{n^2}\sum_{j>m} \mathcal{E}x_{ji}x_{mh}$$

$$= \frac{1}{n}\left\{\mu_\delta + \sum_\gamma \frac{n-\gamma}{n}\left[\mu_{\gamma k+\delta} + \mu_{\gamma k-\delta}\right]\right\}.$$

Thus (4.5) is proved. Then (4.6) is a special case.

We return to the proof of Theorem 2. Now, putting

$$\Delta = \mathcal{E}\sigma_{sy}^2 - \mathcal{E}\sigma_c^2$$

it follows that

$$\Delta = \frac{1}{k}\sum_i \mathcal{E}\bar{x}_i^2 - \mathcal{E}\bar{x}_c^2 + 2\mathcal{E}\bar{x}_c\bar{x} - 2\mathcal{E}\bar{x}^2.$$

Since, from Lemma 2, $\mathcal{E}\bar{x}_i^2$ is independent of i, it follows that

$$\Delta = 2\,\mathcal{E}\,\bar{x}_c\bar{x} - 2\,\mathcal{E}\bar{x}^2.$$

Now, from (4.4) and (4.5), taking $i = c$ and averaging over h, it follows that

$$\mathcal{E}\bar{x}_c\bar{x} = \frac{1}{nk}\left\{\mu_0 + 2\sum_\gamma \frac{n-\gamma}{n}\mu_{\gamma k}\right\} + \frac{2}{nk}\left\{\sum_\epsilon \mu_\epsilon + \sum_\gamma \frac{n-\gamma}{n}\left[\mu_{\gamma k+\epsilon} + \mu_{\gamma k-\epsilon}\right]\right\}.$$

Also,

$$\mathcal{E}\bar{x}^2 = \frac{1}{k^2}\sum_{i,h} \mathcal{E}\bar{x}_i\bar{x}_h,$$

and, since by Lemma 2, $\mathcal{E}\bar{x}_i\bar{x}_h$ depends only on $|\,i-h\,|$, it follows from Lemma 1, that

$$\mathcal{E}\bar{x}^2 = \frac{1}{nk}\left\{\mu_0 + 2\sum_\gamma \frac{n-\gamma}{n}\mu_{\gamma k}\right\}$$

$$+ \frac{2}{nk}\left\{\sum_\delta \frac{k-\delta}{k}\left[\mu_\delta + \sum_\gamma \frac{n-\gamma}{n}\left(\mu_{\gamma k+\delta} + \mu_{\gamma k-\delta}\right)\right]\right\}.$$

Then

$$\Delta = \frac{4}{nk}\sum_\epsilon \frac{\epsilon}{k}\left[\mu_\epsilon - \mu_{k-\epsilon}\right]$$

$$+ \frac{4}{nk}\sum_\gamma \frac{n-\gamma}{n}\sum_\epsilon \frac{\epsilon}{k}\left[\mu_{\gamma k+\epsilon} + \mu_{\gamma k-\epsilon} - \mu_{(\gamma+1)k-\epsilon} - \mu_{(\gamma-1)k+\epsilon}\right].$$

Now

$$\sum_\gamma (n-\gamma)\left[\mu_{\gamma k+\epsilon} - \mu_{(\gamma-1)k+\epsilon}\right] = -n\mu_\epsilon + \sum_{\gamma=0}^{n-1} \mu_{\gamma k+\epsilon},$$

and

$$\sum_\gamma (n-\gamma)\left[\mu_{\gamma k-\epsilon} - \mu_{(\gamma+1)k-\epsilon}\right] = n\mu_{k-\epsilon} - \sum_{\gamma=0}^{n-1} \mu_{(\gamma+1)k-\epsilon}.$$

Hence

$$\sum_{\gamma} \frac{n-\gamma}{n} \sum_{\epsilon} \frac{\epsilon}{k} [\mu_{\gamma k+\epsilon} + \mu_{\gamma k-\epsilon} - \mu_{(\gamma+1)k-\epsilon} - \mu_{(\gamma-1)k+\epsilon}]$$

$$= -\sum_{\epsilon} \frac{\epsilon}{k} (\mu_{\epsilon} - \mu_{k-\epsilon}) + \frac{1}{n} \sum_{\epsilon} \frac{\epsilon}{k} \sum_{\gamma=0}^{n-1} (\mu_{\gamma k+\epsilon} - \mu_{(\gamma+1)k-\epsilon}).$$

Thus, (4.2) is proved. Since $1 \leqq \epsilon \leqq (k-1)/2$, it follows that $\mu_{\gamma k+\epsilon} \geqq \mu_{(\gamma+1)k-\epsilon}$ if $\mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_{nk-1}$. Hence $\Delta \geqq 0$ if the correlogram is monotone decreasing and $\Delta > 0$ if the correlogram is monotone decreasing and not constant.

5. Comments. It is easy to extend the results of this paper to two-dimensional statistical sampling and to the sampling of clusters. These topics will be discussed in following papers.

It is interesting to note that if $\mathcal{E}x^2_{i+(j-1)k}$ is not assumed to be independent of j and i then the above results will not hold without further assumptions concerning the terms $\mathcal{E}x^2_{i+(j-1)k}$.

REFERENCES

[1] W. G. Cochran, "Relative accuracy of systematic and stratified random samples for a certain class of populations," *Ann. Math. Stat.*, Vol. 17 (1946), pp. 164–177.

[2] H. Hotelling and L. M. Solomons, "Limits of a measure of skewness," *Ann. Math. Stat.*, Vol. 3 (1932), pp. 141–142.

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

[4] W. G. Madow, "On the theory of systematic sampling, II," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 333–354.

[5] F. Yates, "Systematic sampling," *Philos. Trans. Roy. Soc. London. Ser. A.*, Vol. 241 (1948), pp. 345–377.