ON THE THEORY OF SYSTEMATIC SAMPLING, I

By William G. Madow and Lillian H. Madow[1,2]

1. Introduction. It is no longer necessary to demonstrate a need for the theory of designing samples. Many of the policy and operating decisions of both government and private industry are based on samples. There has been an increasing tendency in government and industry to make use of sampling theory.[3]

Unfortunately there are still considerable differences between the theory and practice of sampling. The origins of these differences are, on the one hand, the ignorance of administrators concerning the practical contributions that sampling theory can make, and on the other, the lack of sampling theory permitting the evaluation of some useful sampling designs.

Much has been and is being done towards bringing theory and practice into agreement.[4] Administrators and samplers are each successfully educating the others. However, there still exist sampling designs for which an adequate theory has not been developed, even though experience indicates that if such a theory were developed it would demonstrate the superiority of those designs over others for which a theory has been developed.

Perhaps the major omission of sampling theory today is the lack of any statistical method for reaching a decision on whether to take a completely random sample of n elements of a population of N elements, or to take a systematic sample, that is, to begin with element i, and select elements $i, i + k, \cdots, i + (n - 1)k$, as the sample, the starting point i being chosen at random and $N = kn$ approximately.[5] It is with respect to this question of whether to take a systematic[6] or random sample that the statistician is in a dilemma because he has the alternative of recommending a systematic sampling procedure for which no theory exists, or a random sampling procedure that may well yield worse

[1] Bureau of Agricultural Economics and Food Distribution Administration, U. S. Department of Agriculture, Washington, D. C.

[2] Presented at a meeting of the seminar in statistics of the Graduate School, U. S. Department of Agriculture, November 2, 1943.

[3] The recognition of the need for statisticians who know sampling theory has resulted in courses in sampling being given in some of the colleges and universities.

[4] One need only refer to the recent development of positions, the duties of which include giving advice on sampling techniques as well as working in the field of application.

[5] In this paper we will assume that $N = kn$. To do away with that assumption would not add much in the way of generality while it would require some fairly detailed discussion. It may be remarked that when N is not exactly kn, then systematic sampling procedures in which all starting points have equal probability of selection are biased, although the bias is usually trivial. If N is known this bias can be removed by sampling proportionate to possible size of systematic sample.

[6] As we define systematic sampling procedures, a systematic sampling procedure is a random sampling procedure in which many of the C_n^N selections of n from N items are excluded.

results than the systematic procedure. The purpose of this paper is to resolve that conflict by providing an adequate theory of systematic sampling.

In the following sections we present the first parts of our research in the theory of systematic samples. Although this research covers both the theory of sampling single elements and sampling clusters of elements, we shall consider, in this paper, the sampling units to be single elements, not clusters of elements. The latter problem will be dealt with in a later paper. We shall present the theory of systematic sampling both from an unstratified population and a stratified population. Formulas for the mean value and variances of the estimates are derived. Comparisons with random and stratified random sampling designs are made. Furthermore, the estimates of the variances and formulas for "optimum" size and allocation of samples are derived.

A fundamental part of the analysis is the demonstration that from a knowledge of the variance of the population[7] and certain serial correlations or serial variances, can be estimated the variance of estimates based on systematic samples. The basic results are:

a. if the serial correlations have a positive sum, systematic sampling is worse than random sampling,
b. if the serial correlations have a sum that is approximately zero, systematic sampling is approximately equivalent to random sampling, and
c. if the serial correlations have a negative sum, systematic sampling is better than random sampling.

2. The use of a finite population. In this paper we assume, for the calculation of the expected values, that we are sampling from a finite population of elements even though the size of the population may be large enough to permit the use of limiting distributions. Often, this is, mathematically, a matter of choice. The same results would be obtained by assuming a correctly defined multivariate normal distribution and using the notions of conditional probability. From a physical point of view, however, there are several factors that lead to the use of the finite population. We are most frequently sampling an existing population whose laws of transformation are either unknown or not mathematically expressed.[8] Consequently, the notion of a normal or other specified distribution from which we sample and use conditional probability is not part of our thinking concerning the physical problem. On the other hand, if we consider the population to be a finite population, and use a table of random numbers to draw our sample from the finite population, we are using only mathematics implicit in our physical problem. Furthermore, we do obtain a repeatable experiment, that of selecting a random number, that we know is in a state of statistical control.

In the usual problems of the theory of random sampling, the number of

[7] By "variance of population" without further qualification is meant the variance of a random sample of one element of the population.

[8] In other words, our population is not in a state of statistical control over time.

possible samples yielding different sample means is large enough so that the sample means may, with a sufficiently large size of population and sample, be expected to be approximately normally distributed. In systematic sampling, however, the number of possible sample means is usually very small and even if the sizes of population and sample are large, it is difficult to assume a normal distribution. Consequently, in our interpretation of the means and variances of systematic samples we are led to regard the elements of our populations as being the results of single observations on random variables, the distributions of which may vary from element to element. The interpretations that we then make become interpretations of conditional probability, and if the sizes of population and sample are sufficiently large, we can assume that the arithmetic mean of each of the possible sample means is normally distributed.

The theory of systematic sampling under the assumption of an appropriate normal multivariate distribution will be dealt with at a later time.

3. Definitions. Let the finite population to be sampled consist of N elements, $x_1, \cdots, x_N$.

By a sample design is meant the combination of a method of classifying these N elements into k classes that may or may not overlap, and a method of selecting one of these k classes, each class having a designated probability of being selected. The sampling procedure associated with a given sample design is the operation of selecting one of the k classes according to the method stated in the sample design. The sample is the particular class obtained by the sampling procedure.

By a random sampling procedure is meant any sampling procedure such that if the sampling design yields k classes then the probability of selecting anyone of these classes is $1/k$. Any sample design having a random sampling procedure associated with it is a random sampling design. One of the nonrandom sampling procedures that is being used is the procedure in which the classes have associated to them numbers, called sizes, and the probability of a given class being the sample is proportionate to its size.[9] Other nonrandom sampling procedures are doubtless being used.

By an unrestricted random sampling design for selecting n elements from N elements is meant the sampling design such that there are C_n^N classes, the possible selections of n from N elements, each having a probability of $1/C_n^N$ of being the sample. The associated random sampling procedure might consist in identifying each class by a number i, $i = 1, \cdots, C_n^N$ and selecting a number i from a table of random numbers. The random sampling procedure might also consist in identifying the N elements with numbers $j = 1, \cdots, N$, and then selecting a number j from a table of random numbers, then selecting a different number j from a table of random numbers, and following that procedure until n numbers

[9] For a discussion of this problem see the paper entitled, "On the theory of sampling from finite populations," by Morris H. Hansen and William N. Hurwitz, *Annals of Math. Stat.*, Vol. 14 (1943), pp. 333-362.

from 1, $\cdots$, N without repetition have been selected from the table of random numbers. The elements associated with these integers would be a random sample. It is easy to see that the two procedures are equivalent.

A random sampling design that is not unrestricted is said to be restricted. There are many types of restricted random sampling designs of which what we call systematic designs are only one. Among these restricted designs are stratified, cluster, double, matched, polynomial, and other sampling designs, each having been developed as attempts to bring theory and practice together, to suggest improvements in practice, and to solve problems arising in practice.

By a systematic sampling design is meant a classification of the N elements into k classes, $S_1, \cdots, S_k$ where S_i consists of $x_i, x_{i+k}, \cdots, x_{i+(n-1)k}$, and a random sampling procedure for selecting one of the S_i.

It is thus clear that a systematic sampling design is a type of cluster sampling design. It will be shown that the new aspect of cluster sampling introduced in systematic sampling is that a knowledge of the order of the elements in the population is used to obtain the values of the intraclass correlation coefficient and changes in the value of that coefficient as the size of sample changes.

Sampling designs may involve combinations of random and systematic sampling designs, as well as random and nonrandom sampling procedures.

The population from which these samples are drawn may or may not be stratified and the sampling units may be single elements or clusters of elements.

4. Bases for selecting among sample designs. From the many sampling designs that can be constructed in order to obtain desired estimates, one will be chosen for use on the bases of administrative considerations, cost, and sampling error. It has become customary, on the basis of limiting distribution theory and the theory of best linear unbiased estimates to use the standard deviation of the sample estimate about the character estimated as the measure of sampling error.

Although in this paper we shall continue this practice, it must be pointed out that as more sampling designs are constructed, there is the danger that for some of these designs the limiting distribution theory is not valid, and the use of the standard error becomes more a matter of custom than the result of analysis. This danger is present for systematic sampling designs and is being further investigated.

It is perhaps desirable to remark that bias, consistency, and efficiency are properties of the sampling design and estimation functions used, not of the particular sample obtained. Any estimate based on a sample will probably differ from the character estimated. It is the function of statistical analysis to indicate how large this difference may be.

5. Notation. The letter, P, with appropriate subscripts is used for population, and subpopulations such as strata.

The number of strata is denoted by L, and the number of elements in the i^{th} stratum is denoted by N_i. Sizes of sample are denoted by n with appropriate subscripts.

The arithmetic mean of the elements of a population or subpopulation is denoted by $\bar{x}$ with appropriate subscripts.

Any particular subclass of a population as defined by the sampling design is denoted by S with subscripts. Estimates based on an S with subscripts are denoted by $\tilde{x}$ with subscripts.

6. Unstratified systematic sampling, the sampling unit consisting of one element. The values assumed by the subscripts used in this section are given in Appendix A.

Let the population, P, consist of N elements $x_1, \cdots, x_N$. It is desired to estimate the arithmetic mean, $\bar{x}$, of P.

Let[5] $N = kn$, and let the class S_i consist of the n elements $x_i, x_{i+k}, \cdots x_{i+(n-1)k}$. Then, the systematic sampling design for estimating $\bar{x}$ from a sample of size n, consists of the k classes, $S_1, \cdots, S_k$, and the requirement that the sampling procedure be such that the probability is $1/k$, that S_i is the class selected by the sampling procedure.

Let $\bar{x}_i$ be the arithmetic mean of the elements of S_i, i.e., $n\bar{x}_i = x_i + x_{i+k} + \cdots + x_{i+(n-1)k}$, and let $\tilde{x}$ be the sample mean, i.e., $\tilde{x} = \bar{x}_i$ if S_i is selected by the sampling procedure.

In dealing with systematic sampling, we shall have occasion to use both the circular and non-circular definitions of the serial correlation coefficients and the associated serial variances.

We shall assume that if $h > kn$ then $x_h = x_{h-kn}$. This is used in the circular definitions.

Let
$$kn\sigma^2 = \sum_{\nu}(x_\nu - \bar{x})^2,$$
and let
$$knC_{k\mu} = \sum_{\nu}(x_\nu - \bar{x})(x_{\nu+k\mu} - \bar{x}).$$

Then, the circular definition of the serial correlation coefficient $\rho_{k\mu}$ is $\sigma^2\rho_{k\mu} = C_{k\mu}$, which we shall use unless n is even, when we define $\rho_{kn/2}$ by the equation

$$2\sigma^2\rho_{kn/2} = C_{kn/2},$$

in order to simplify the writing of the formula for $\sigma_{\tilde{x}}^2$.

Similarly, if we define the serial variance, $s_{k\mu}$, by the equation $kns_{k\mu} = \sum_{\nu} (x_\nu - x_{\nu+k\mu})^2$, then we are using the circular definition of the serial variance. The circular definition of the serial variance ratio $v_{k\mu}$ is then $\sigma^2 v_{k\mu} = s_{k\mu}$ which we shall use unless n is even, when we define $v_{kn/2}$ by the equation

$$2\sigma^2 v_{kn/2} = s_{kn/2}.$$

The non-circular definitions of the serial correlations and serial variances are given by

$$(1)\qquad \begin{aligned} k(n-\delta)C'_{k\delta} &= \sum_j (x_j - \bar{x})(x_{j+k\delta} - \bar{x}),\\ \sigma^2\rho'_{k\delta} &= C'_{k\delta},\\ k(n-\delta)s'_{k\delta} &= \sum_j (x_j - x_{j+k\delta})^2,\end{aligned}$$

and

$$\sigma^2 v'_{k\delta} = s'_{k\delta}.$$

The intraclass correlation coefficient $\bar{\rho}_k$ is defined by the equation

$$\sigma^2\bar{\rho}_k = \mathcal{E}(x_\mu - \bar{x})(x_\nu - \bar{x}),$$

where the random process consists in first sampling one of the S_i at random and then selecting two of the x's at random from the S_i that was selected. Then, since

$$k\sigma^2_{\tilde{\bar{x}}} = \sum_i (\bar{x}_i - \bar{x})^2,$$

and,

$$(2)\qquad \sigma^2\bar{\rho}_k = (n/n-1)\sigma^2_{\tilde{\bar{x}}} - (1/n-1)\sigma^2$$

we have

$$(3)\qquad \sigma^2_{\tilde{\bar{x}}} = \frac{1}{n}(1 + (n-1)\bar{\rho}_k)$$

It is easy to see from (1) that the intraclass correlation coefficient is given by

$$\begin{aligned}\bar{\rho}_k &= \frac{2}{n(n-1)}\sum_\delta (n-\delta)\rho'_{k\delta}\\ &= \frac{2}{n-1}\sum_\mu \rho_{k\mu},\end{aligned}$$

and that consequently, if n is odd, $\bar{\rho}_k$ is the arithmetic mean of the $\rho_{k\mu}$ while if n is even, $\bar{\rho}_k$ is equal to the arithmetic mean of the $\rho_{k\mu}$ multiplied by $n/(n-1)$.

THEOREM[10]: *Using the systematic sampling design, the estimate $\tilde{\bar{x}}$ is an unbiased estimate of $\bar{x}$, and has variance $\sigma^2_{\tilde{\bar{x}}}$ where*

$$(4)\qquad \begin{aligned}\sigma^2_{\tilde{\bar{x}}} &= \sigma^2\left\{1 - \frac{1}{n^2}\sum_\delta (n-\delta)v'_{k\delta}\right\}\\ &= \sigma^2\left(1 - \frac{1}{n}\sum_\mu v_{k\mu}\right)\\ &= \frac{\sigma^2}{n}\left\{1 + \frac{2}{n}\sum_\delta (n-\delta)\rho'_{k\delta}\right\}\\ &= \frac{\sigma^2}{n}\left(1 + 2\sum_\mu \rho_{k\mu}\right)\\ &= \frac{\sigma^2}{n}\{1 + (n-1)\bar{\rho}_k\}.\end{aligned}$$

[10] A proof of Theorem 1 that is somewhat simpler to follow but which, in the authors opinion, is not as informative as that given below could be obtained by substituting for $\bar{\rho}_k$ using equations (2) and (3). The lemmas in Appendix B are, of course, of interest in themselves in finite sampling.

PROOF: From the definitions of expected value, $\bar{x}_i$, $\tilde{x}$, and the systematic sampling design, it follows that $\tilde{x}$ is a variate with possible values $\bar{x}_1, \cdots, \bar{x}_k$, the probability that $\tilde{x} = \bar{x}_i$ being $1/k$. Then

$$k\mathcal{E}\tilde{x} = \bar{x}_1 + \cdots + \bar{x}_k, \tag{5}$$

and, when the values of the $\bar{x}_i$ are substituted in (5), it follows that $\mathcal{E}\tilde{x} = \bar{x}$, that is, $\tilde{x}$ is an unbiased estimate of $\bar{x}$.

Having calculated $\mathcal{E}\tilde{x}$, it is necessary to calculate $\mathcal{E}\tilde{x}^2$ in order to evaluate $\sigma^2_{\tilde{x}}$. From the definition of expected values, it follows that

$$k\mathcal{E}\tilde{x}^2 = \bar{x}_1^2 + \cdots + \bar{x}_k^2, \tag{6}$$

and when the values of the $\bar{x}_i$ are substituted in (6), it follows that

$$n^2k\mathcal{E}\tilde{x}^2 = \sum_{i,\alpha,\gamma} x_{i+(\alpha-1)k}\, x_{i+(\gamma-1)k} \tag{7}$$

Then, when $f(u)$ is replaced by u in Lemma 6 of Appendix B, it follows, from the definition of the variance, that $\sigma^2_{\tilde{x}} = \left(\frac{1}{kn}\right) \sum_{\nu} (x_\nu - \bar{x})^2 - \left(\frac{1}{kn^2}\right) \sum_{\delta,j} (x_j - x_{j+k\delta})^2 = \sigma^2 - \frac{1}{n^2} \sum_{\delta} (n - \delta) s'_{k\delta}$, and when $f(u)$ is replaced by u in Lemma 8, it follows that

$$\begin{aligned} \sigma^2_{\tilde{x}} &= \left(\frac{1}{kn^2}\right) \sum_{\nu} (x_\nu - \bar{x})^2 + \left(\frac{2}{kn^2}\right) \sum_{\delta,j} (x_j - \bar{x})(x_{j+k\delta} - \bar{x}) \\ &= \left(\frac{1}{n}\right) \sigma^2 + \frac{2}{n^2} \sum_{\delta} (n - \delta) c'_{k\delta}. \end{aligned}$$

If in Lemma 9 of Appendix B we now replace $f(x_j, x_{j+k\delta})$ by $(x_j - x_{j+k\delta})^2$ then $\sigma^2_{\tilde{x}} = \sigma^2 - \left(\frac{1}{n}\right) \sum_{\mu} s_{k\mu}$,

and if we replace $f(x_j, x_{j+k\delta})$ by $(x_j - \bar{x})(x_{j+k\delta} - \bar{x})$ then

$$\sigma^2_{\tilde{x}} = \frac{1}{n} (\sigma^2 + 2 \sum_{\mu} c_{k\mu}).$$

Finally, we have, then

$$\sigma^2_{\tilde{x}} = \sigma^2 \left(1 - \frac{1}{n} \sum_{\mu} v_{k\mu}\right),$$

and

$$\sigma^2_{\tilde{x}} = \frac{\sigma^2}{n} (1 + 2 \sum_{\mu} \rho_{k\mu}).$$

7. Possible values of the $\rho'_{k\delta}$, $\rho_{k\mu}$ and $\sigma^2_{\tilde{x}}$. Let us investigate briefly the effects of different patterns of variation on the values of $\rho'_{k\delta}$ and $\sigma^2_{\tilde{x}}$. Now $\sigma^2\rho'_{k\delta} = \frac{1}{k(n - \delta)} \sum_{j} (x_j - \bar{x})(x_{j+k\delta} - \bar{x})$. Suppose that $x_i = x_{i+k\delta}$, $\delta = 1, \cdots, n - 1$,

$i = 1, \cdots, k$. Then $\sum_{\nu} (x_\nu - \bar{x})^2 = n \sum_{i} (x_i - \bar{x})^2$, and $\sum_{j} (x_j - x)(x_{j+k\delta} - \bar{x}) = (n - \delta) \sum_{i} (x_i - \bar{x})^2$. Upon substitution it follows that $\rho'_{k\delta} = 1$, and $\sigma^2_{\bar{x}} = \sigma^2$. This result for $\sigma^2_{\bar{x}}$ is intuitively clear, since all the variability is among the possible samples, and thus any particular systematic sample is equivalent to one observation.

Suppose, on the other hand, that $x_{k\delta+\alpha} = x_{k\delta+\beta}\,\alpha, \beta = 1, \cdots, k; \delta = 1, \cdots, n - 1$. Then $\sum_{\nu} (x_\nu - \bar{x})^2 = k \sum_{\alpha} (x_{i+(\alpha-1)k} - \bar{x})^2$ for any i, $i = 1, \cdots, k$, and $\sum_{j} (x_j - \bar{x})(x_{j+k\delta} - \bar{x}) = k \sum_{\lambda} (x_{i+(\lambda-1)k} - \bar{x})(x_{i+(\lambda+\delta-1)k} - \bar{x})$. Furthermore $0 = [\sum_{\alpha} (x_{i+(\alpha-1)k} - \bar{x})]^2 = \sum_{\alpha} (x_{i+(\alpha-1)k} - \bar{x})^2 + 2 \sum_{\lambda,\delta} (x_{i+(\lambda-1)k} - \bar{x})(x_{i+(\lambda+\delta-1)k} - \bar{x})$. Hence

$$2\sum_{\delta} \frac{n - \delta}{n} \rho'_{k\delta} = -1 \quad \text{and} \quad \sigma^2_{\bar{x}} = 0.$$

It is possible to construct examples in which any particular $\rho'_{k\delta} = -1$, but in such cases the remaining $\rho'_{k\delta}$ each vanish. It is well known that the minimum value of $\bar{\rho}_k$ is $-1/(n - 1)$.

Finally, let us consider the expected values of $\rho'_{k\delta}$ and $\sigma^2_{\bar{x}}$ if the x's have been assigned their subscripts at random. These values are $\rho'_{k\delta} = -1/(nk - 1)$ and $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{nk - n}{nk - 1}\right)$.

In most practical applications of systematic sampling it will be highly unlikely that the distribution of the x's will be such that the x's may be said to have been assigned their subscripts at random. In general, there will be logical reasons to expect that the x's will have some fundamental trend. Thus, information will often be available, or may be obtained by a small subsample, on the basis of which a decision can be made to use some approach differing from that of assuming the subscripts of the x's to have been assigned at random.

8. Estimates of the parameters. The formulae obtained in section 6 for the variance of the mean of a systematic sample are population formulae. Their values depend on the values of all the elements of the population. However, even in tests of possible sampling procedures, we rarely have available the resources with which to study the entire population. Consequently, it becomes necessary to investigate the possibility of estimating the population variances and serial correlations from samples. It will be shown that the estimates of the variances and correlations derived from a single S_i are biased and inconsistent whereas it will be possible to construct unbiased or consistent estimates from samples of more than one of the S_i. The sampling variations of these estimates must be left for further study.

Let us assume that instead of sampling only one of the S_i, as we did in section

6, we sampled g of the S_i at random. Then our sample would consist of all the elements in the S_β. The sample mean, $\hat{x}$, is defined by

$$g\hat{x} = \sum_\beta \bar{x}_\beta$$

if the subscripts of our sample classes are $i_1, \cdots, i_g$.

Then it is easy to see that $\hat{x}$ is unbiased Furthermore, since we can regard this sampling procedure as the sampling of g of k elements at random, it follows that $\sigma_{\hat{x}}^2 = \frac{k-g}{k-1}\frac{1}{g}\,\sigma_{\bar{x}}^2$ and, we have evaluated $\sigma_{\bar{x}}^2$ in section 6.[11]

Since

$$k\sigma_{\bar{x}}^2 = \sum_i (\bar{x}_i - \bar{x})^2,$$

we shall consider estimating $\sigma_{\bar{x}}^2$ by $s_{\hat{g}}^2$ where

$$gs_{\hat{g}}^2 = \sum_\beta (\bar{x}_\beta - \hat{x})^2.$$

Now, since $E\,\hat{x}^2 = \sigma_{\hat{x}}^2 + \bar{x}^2$, and

$$E\sum_\beta \bar{x}_\beta^2 = \frac{g}{k}\sum_i \bar{x}_i^2 = g(\sigma_{\bar{x}}^2 + \bar{x}^2)$$

it follows that $\mathcal{E}s_{\hat{g}}^2 = \sigma_{\bar{x}}^2$, and hence $s_{\hat{g}}^2$ is an unbiased estimate of $\sigma_{\bar{x}}^2$. Furthermore $\mathcal{E}s_{\hat{g}}^2 = \frac{g(k-1)}{k-g}\,\sigma_{\hat{x}}^2$.

We now turn to estimates of the $\rho_{k\mu}$ and σ^2.

Let

$$gn\hat{s}_g^2 = \sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \hat{x})^2$$

and let

$$gn\hat{c}_{k\mu g} = \sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \hat{x})(x_{\beta+(\alpha+\mu-1)k} - \hat{x}).$$

Then, it may be shown that

$$\mathcal{E}\hat{s}_g^2 = \sigma^2 - \sigma_{\hat{x}}^2,$$

and

$$\mathcal{E}\hat{c}_{k\mu\delta} = C_{k\mu} - \sigma_{\hat{x}}^2.$$

Hence

$$\mathcal{E}\hat{s}_g^2 + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right) = \sigma^2$$

[11] It may be wondered why we sample these S_i at random rather than systematically. If we sampled the S_i systematically, it would be equivalent to taking a single systematic sample having smaller intervals between elements of the sample. Furthermore, we could not derive the unbiased estimates of the sampling variance that we can now.

and

$$\mathcal{E}\hat{c}_{\lambda\mu g} + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right) = C_{k\mu}\,.$$

The estimate, $r_{k\mu}$, of $\rho_{k\mu}$ defined by

$$\hat{c}_{k\mu g} + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right) = r_{k\mu}\left[\hat{s}_g^2 + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right)\right]$$

is thus biased but in many cases the bias will be small Of course, if $\mu = \frac{n}{2}$ when n is even then $r_{kn/2}$ is multiplied by 2 to estimate $\rho_{kn/2}$ as previously defined. Another approach would be to consider

$$gn\,{}_w\hat{s}_g^2 = \sum_{\beta,\alpha}(x_{\beta+(\alpha-1)k} - \bar{x}_\beta)^2,$$

and

$$gn\,{}_w\hat{c}_{k\mu g} = \sum_{\beta,\alpha}(x_{\beta+(\alpha-1)k} - \bar{x}_\beta)(x_{\beta+(\alpha+\mu-1)k} - \bar{x}_\beta).$$

When that is done, it follows that

$$\mathcal{E}\,{}_w\hat{s}_g^2 = \sigma^2 - \sigma_{\bar{\bar{x}}}^2\,,$$

and

$$\mathcal{E}\,{}_w\hat{c}_{k\mu g} = C_{k\mu} - \sigma_{\bar{\bar{x}}}^2\,,$$

and

$$\mathcal{E}({}_w\hat{s}_g^2 + s_{\hat{g}}^2) = \sigma^2,$$

$$\mathcal{E}({}_w\hat{c}_{k\mu g} + s_{\hat{g}}^2) = C_{k\mu}\,.$$

Another estimate of $\rho_{k\mu}$ is thus defined by the equation

$${}_w\hat{c}_{k\mu g} + s_{\hat{g}}^2 = {}_wr_{k\mu}[{}_w\hat{s}_g^2 + s_{\hat{g}}].$$

When $g = 1$, $s_{\hat{g}}^2 = 0$ and we are unable to provide unbiased estimates of σ^2, $C_{k\mu}$, and $\sigma_{\bar{\bar{x}}}^2$ from the sample. However, since

$$\frac{1 - r_{k\mu}}{1 - r_{k\mu'}} = \frac{\hat{s}_g^2 - \hat{c}_{k\mu g}}{\hat{s}_g^2 - \hat{c}_{k\mu' g}},$$

it follows that approximately

$$\frac{1 - \rho_{k\mu}}{1 - \rho_{k\mu'}} = \mathcal{E}\,\frac{1 - r_{k\mu}}{1 - r_{k\mu'}},$$

since $\mathcal{E}[\hat{s}_g^2 - \hat{c}_{k\mu g}] = \sigma^2 - C_{k\mu}$. Similar equations hold for the ${}_wr_{k\mu}$.

When we estimate the $\rho'_{k\delta}$, then the "within class" definition is simpler. Let $k(n - \delta)_w c'_{k\delta g} = \sum_{i,\lambda} (x_{i+(\lambda-1)k} - {}_{1\delta}\bar{x}_i)(x_{i+(\lambda+\delta-1)k} - {}_{2\delta}\bar{x}_i)$, where

$$(n - \delta)_{1\delta}\bar{x}_i = \sum_{\lambda} x_{i+(\lambda-1)k} ,$$

$$(n - \delta)_{2\delta}\bar{x}_i = \sum_{\lambda} x_{i+(\lambda+\delta-1)k} ,$$

and let

$$g(n - \delta)_w \hat{c}'_{k\delta g} = \sum_{\beta,\lambda} (x_{\beta+(\lambda-1)k} - {}_{1\delta}\bar{x}_\beta)(x_{\beta+(\lambda+\delta-1)k} - {}_{2\delta}\bar{x}_\beta).$$

Let

$$k_b c'_{k\delta g} = \sum_{i} ({}_{1\delta}\bar{x}_i - \bar{x})({}_{2\delta}\bar{x}_i - \bar{x}),$$

and let

$$g_b \hat{c}_{k\delta g} = \sum_{\beta} ({}_{1\delta}\bar{x}_\beta - \bar{x})({}_{2\delta}\bar{x}_\beta - \bar{x}),$$

Then

$$c'_{k\delta} = {}_w c'_{k\delta g} + {}_b c'_{k\delta g} ,$$

and

$$\mathcal{E}({}_w\hat{c}'_{k\delta g} + {}_b\hat{c}'_{k\delta g}) = C'_{k\delta} .$$

Thus, as estimates of the $\rho'_{k\delta}$ we obtain $r'_{k\delta}$ where

$${}_w\hat{c}'_{k\delta g} + {}_b\hat{c}'_{k\delta g} = r'_{k\delta}({}_w s_g^2 + s_g^2).$$

In cases where $\bar{x}$ is known simpler estimates of the $\rho_{k\mu}$, $\rho'_{k\delta}$, and σ^2 may be easily obtained since

$$\mathcal{E}\sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \bar{x})(x_{\beta+(\alpha+\mu-1)k} - \bar{x}) = gnC_{k\mu} ,$$

$$\mathcal{E}\sum_{\beta,\lambda} (x_{\beta+(\lambda-1)k} - \bar{x})(x_{\beta+(\lambda+\delta-1)k} - \bar{x}) = g(n-\delta)C'_{k\delta} ,$$

and

$$\mathcal{E}\sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \bar{x})^2 = gn\sigma^2.$$

Thus, in pilot studies, when $\bar{x}$ is known it is possible to estimate the parameters in $\sigma_{\bar{x}}^2$ from even a single sample.

9. Changes in the variance with changing size of sample. The chief reasons for expressing the variance of a systematic sampling design in terms of the variance of a random sample and the serial correlation coefficients were

1. To enable the making of comparisons with random and other sampling designs

2. To simplify the analysis of causes for the difference in the efficiencies of the systematic and random designs, and
3. To simplify the making of estimates of the variance for different sizes of sample.

In this section we are concerned with the third of these reasons. We shall discuss only the $\rho_{k\mu}$ since the analysis in terms of the $\rho'_{k\delta}$ is very similar.

The problem with which we are concerned is the estimation of the function, $\bar{\rho}_k$, of k. In order to show how this may be done for all values of k when the $\rho_{k\mu}$ have been computed for one value of k, let us first note that since σ^2 does not depend on k we may confine our considerations to the $C_{k\mu}$. In section 6 we have defined $C_{k\mu}$ by the equation

$$knC_{k\mu} = \sum_{\nu} (x_\nu - \bar{x})(x_{\nu+k\mu} - \bar{x}).$$

Thus, if we wish to evaluate the $C_{k'\mu'}$ where k' is such that $k' \neq k$ and $k'n' = kn = N$, we have the result $C_{k'\mu'} = C_{k\mu}$ if $k'\mu' = k\mu$ and, thus, for any given values k' and μ', we have

$$C_{k'\mu'} = C_{k\ k'\mu'/k}$$

where we have replaced μ by $\dfrac{k'\mu'}{k}$.

This procedure will involve, if $k' < k$, some interpolation, but if the $\rho_{k\mu}$ are plotted against μ, this interpolation may often be carried through graphically. However, it is usually advisable to take k so that the possible values of k' are such that $k' > k$.

In some cases it may be possible to construct a correlation function. For example, if the x_ν may be represented by a polynomial in ν, then $\rho'_{k\delta}$ may be represented by a polynomial in δ. From that fact we conclude that if the x_ν vary about a smooth trend the $\rho'_{k\delta}$ will also vary about a smooth trend and it may be possible to interpolate. Further investigation of this problem is necessary.

10. Stratified systematic sampling. In sampling practice it is customary to deal with stratified populations. The variance of an estimate based on a stratified population will usually not include the variability among the strata. Consequently, when a population is well stratified the variability of estimates based in a sample of size n will usually be considerably less than the variability of an estimate based on a random sample of size n, ignoring the strata. We now discuss the theory of systematic sampling from a stratified population.

Let us assume that the population, P, consists of L strata, $P_1, \cdots, P_L$, the αth of which contains N_α elements $x_{\alpha 1}, \cdots, x_{\alpha N_\alpha}$. It is desired to estimate the arithmetic mean, $\bar{x}$ of P. Let the arithmetic mean of P_α be denoted by $\bar{x}_\alpha$. Let $N_\alpha = k_\alpha n_\alpha$.

We shall consider two possible cases, the first of which is often used because of the administrative simplicity of giving identical operating instructions to the people selecting samples in different places. The results of this section will indicate when this method may be used.

Sampling Procedure I—Suppose that $k_1 = k_2 = \cdots k_L = k$, and that the sampling procedure consists in selecting one of the integers, $1, \cdots, k$ at random, each integer having a probability $1/k$ of being selected. Then, if the integer selected is, for example, i, the sample of P_a consists of $x_{ai}, x_{ai+k}, \cdots, x_{ai+(n_a-1)k}$. Thus, there are exactly k possible samples, $S_1, \cdots, S_k$, each having probability $1/k$ of being the actual sample obtained by performing the sampling procedure.

Sampling Procedure II—The sampling procedure consists in selecting one of the integers $1, \cdots, k_a$ at random, for each value of a, each integer having a probability of $1/k_a$ of being selected. Then, there are exactly $k_1 \cdot \cdots \cdot k_L$ possible samples, each having probability $1/k_1 \cdot \cdots \cdot k_L$ of being the actual sample obtained by performing the sampling procedure.

Other sampling procedures for stratified sampling, of course, exist. The two listed above, however, cover most practical problems except those involving cluster sampling. These will be treated in a later paper. Furthermore, from the conclusions derived concerning these procedures it will be possible to infer conclusions concerning other stratified sampling procedures.

Let S_{ai} be the class of elements $x_{ai}, x_{ai+k}, \cdots, x_{ai+(n_a-1)k}$. We consider sampling procedure I. A systematic sample of size n_a is to be selected from P_a. The possible samples are $S_1, \cdots, S_k$ where S_i consists of all the elements in $S_{1i}, \cdots, S_{Li}$. Let the arithmetic mean of the elements in S_{ai} be denoted by $\tilde{x}_{ai}$. Let the arithmetic mean of the sample from P_a be denoted by $\tilde{x}_a$ and let the sample mean be denoted by $\tilde{x}$, where

$$N\tilde{x} = N_1\tilde{x}_1 + \cdots + N_L\tilde{x}_L .$$

Then $N\mathfrak{E}\tilde{x} = \sum_a N_a \mathfrak{E}\tilde{x}_a = \sum_a N_a \frac{1}{k} \sum_i \tilde{x}_{ai} = N\bar{x}.$

It follows from Appendix C, that

$$\sigma_{\tilde{x}}^2 = \frac{1}{N^2} \sum_{a,b} N_a N_b \, \sigma_{\tilde{x}_a \tilde{x}_b}$$

where

$$\sigma_{\tilde{x}_a \tilde{x}_b} = \mathfrak{E}(\tilde{x}_a - \bar{x}_a)(\tilde{x}_b - \bar{x}_b)$$

$$= \frac{1}{k} \sum_i (\tilde{x}_{ai} - \bar{x}_a)(\tilde{x}_{bi} - \bar{x}_b).$$

Although the expression for $\sigma_{\tilde{x}_a \tilde{x}_b}$ can be further simplified, the important fact is that if corresponding items in different strata are positively correlated, it is inadvisable to use sampling procedure I unless other considerations than sampling error are dominant. But if the corresponding items are negatively correlated then sampling procedure I will yield a smaller variance than sampling procedure II.

We now consider sampling procedure II. The difference between sampling procedures I and II is that in sampling procedure II we know that $\sigma_{\bar{x}_a \bar{x}_b} = 0$, if $a \neq b$ because of the separate selection of sample in each stratum. Thus, under sampling procedure II, $\sigma_{\bar{x}}^2 = \frac{1}{N^2} \sum_a N^2 \sigma_{\bar{x}_a}^2$ where $\sigma_{\bar{x}_a}^2$ has been derived in section 6.

11. A comparison of the efficiences of systematic and random sampling procedures. The study of any sampling technique is incomplete unless some comparisons are made with other possible sampling techniques. In this section the systematic sampling procedure is compared with the unrestricted random and stratified random sampling procedure.

The means and variances associated with the random and stratified random sampling procedures will be denoted by the use of primes (′) and double primes (″) respectively.

Then we know that

$$\sigma_{\bar{x}}^2/\sigma_{\bar{x}'}^2 = (1 + 2 \sum_\mu \rho_{k\mu}) \left(\frac{kn - 1}{kn - n}\right)$$

and consequently $\sigma_{\bar{x}}^2 < \sigma_{\bar{x}'}^2$ if

$$\sum_\mu \rho_{k\mu} < -(n - 1)/2(kn - 1).$$

If n is large relative to k, we may use $-\frac{1}{2}k$ as an approximation to $-(n - 1)/2(kn - 1)$.

In order to make more specific comparisons, it is useful to assume that the population elements x_ν are given by some function of ν, and to assume some functions such as

$$x_\nu = A_0 + A_1 \nu + \cdots + A_h \nu^h,$$

or

$$x_\nu = B_0 + A_1 \sin \frac{2\pi\nu}{N} + B_1 \cos \frac{2\pi\nu}{N}$$

$$+ \cdots$$

$$+ A_h \sin \frac{2\pi h\nu}{N} + B_h \cos \frac{2\pi h\nu}{N},$$

and then to investigate the efficiencies of the various possible sampling procedures on the bases of such assumed distributions of the x_ν. It should be noted that the use of the systematic sampling technique involves the assumption that it is possible to order the elements of the population in a logical way, and then use this ordering in selecting the sample systematically.

We shall now consider several possibilities. Let us first note that if we are sampling but one element from a stratum then the variance of the stratum sample mean is the same whether the sampling is random or systematic. On the other hand, it follows from section 10 that if we stratify the population intô L strata so that a systematic sample of size L chooses the jth element of each stratum, say, then the variance of the mean of the stratified random sample will be greater or less than the variance of the mean of the systematic sample depending on whether the average correlation between strata sample means in the systematic sample is negative or positive.

Let us now consider the origin of the warnings against the use of systematic samples from a population having a periodic distribution. If k is the period, the correlation between the strata means of the systematic sample is $+1$ and hence the random sample.is superior. However, if the period is $2k$ then we shall show that the systematic sample will probably have a smaller variance.

Suppose that the period is $2k$ and that within two adjoining strata of size k we always have $x_1 = x_{2k}$, $x_2 = x_{2k-1}$, $\cdots$, $x_k = x_{k+1}$ and $x_i - \bar{x} = -(x_{k+i} - \bar{x})$. Then, if we are sampling one element from each stratum, the correlation between the systematic sample means, (the individual elements in this case), will be -1 if the strata subscripts differ by an odd number and $+1$ if the strata subscripts differ by an even number.

The variance within each of the n strata is σ_1^2, where

$$k\sigma_1^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2.$$

The variance between strata means is zero. Hence $\sigma^2 = \sigma_1^2$. The variance of the mean of an unrestricted random sample of size n where $n = L$ is then $\sigma_{\bar{x}'}^2 = \frac{N - L}{N - 1} \frac{\sigma_1^2}{L}$ and the variance of the stratified random sampling mean is $\sigma_{\bar{x}''}^2 = (1/L)\sigma_1^2$ while the variance of the systematic sampling mean is

$$\sigma_{\bar{x}}^2 = \frac{\sigma_1^2}{L^2} \sum_{i,j=1}^{L} (-1)^{i-j}(2 - \delta_{ij})$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Then it may be shown that if L is even $\sigma^2 = 0$ while, if L is odd $\sigma^2 = (1/L^2)\sigma_1^2$. Consequently, the efficiency of the systematic sample mean is greater than the efficiency of the stratified random sample mean if the population has a periodic distribution and the size of stratum is half the period. It should be noted that the same situation holds for k equal to an even or odd multiple of half the period as held for k equal to the period of half the period.

The situation is quite different if we assume that the elements of the population have a straight line distribution. Without loss of generality, we may assume that the straight line distribution is given by $x_\nu = \nu$. Then for an unrestricted random sample of size n, the sample mean being denoted by $\bar{x}'$ we have $\mathcal{E}\bar{x}' = \bar{x} = \frac{1}{2}(kn + 1)$,

$$\sigma^2 = \frac{k^2 n^2 - 1}{12},$$

and

$$\sigma_{\bar{x}'}^2 = \frac{(k-1)(kn+1)}{12}.$$

For a stratified random sampling design, let us assume that $N_1 = \cdots = N_L = \frac{cN}{n}$ where c may equal any of the integers $1, 2, \cdots, n$; i.e. $L = \frac{n}{c}$. Let $n_1 = \cdots = n_L = c$. Then $\sigma_{\bar{x}''}^2 = \frac{c^2}{n^2}\frac{k-1}{ck-1}\sum_{\alpha}\sigma_\alpha^2$ where $\bar{x}''$ is the sample mean of the stratified random sample. If the αth stratum contains $x_{(\alpha-1)\frac{cN}{n}+1}, \cdots, x_{\alpha\frac{cN}{n}}$ then $\sigma_\alpha^2 = \frac{c^2k^2-1}{12}$ and $\sigma_{\bar{x}''}^2 = \frac{c}{n}\cdot\frac{k-1}{ck-1}\,\frac{c^2k^2-1}{12}$. Finally

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \sigma^2 - \frac{1}{kn^2}\sum_{\delta}\sum_{j}(x_j - x_{j+k\delta})^2\\ &=. \sigma^2 - \frac{k^2(n^2-1)}{12}\\ &= \frac{k^2-1}{12}.\end{aligned}$$

To summarize

$$\sigma_{\bar{x}'}^2 = \frac{(k-1)(kn+1)}{12} = \frac{(k-1)(N+1)}{12},$$

$$\sigma_{\bar{x}''}^2 = \frac{c(k-1)(ck+1)}{12n} = \frac{(k-1)\left(\frac{N}{L}+1\right)}{12L},$$

$$\sigma_{\bar{x}}^2 = \frac{k^2-1}{12}.$$

It is clear that both $\sigma_{\bar{x}}^2$ and $\sigma_{\bar{x}''}^2$ are less than $\sigma_{\bar{x}'}^2$. However $\sigma_{\bar{x}}^2/\sigma_{\bar{x}''}^2 = \frac{L(k+1)}{\frac{N}{L}+1}$.

Since $kn = N$ and $cL = n$ it follows that $k = \frac{N}{cL}$ and hence $\sigma_{\bar{x}}^2 < \sigma_{\bar{x}''}^2$ if $N > L(L-1)$ and $c \geq \frac{L}{1 - \frac{L(L-1)}{N}}$. In all cases $\frac{N}{L} \geq c$. It follows therefore that for a value of c to exist we must have $\frac{N}{L} > \frac{NL}{N - L(L-1)}$ as a result of which we find that N must exceed $2L^2 - L$. Hence $\sigma_{\bar{x}}^2 \leq \sigma_{\bar{x}''}^2$ if $N > 2L^2 - L$ and $c \geq \frac{L}{1 - \frac{L(L-1)}{N}}$. Otherwise $\sigma_{\bar{x}}^2 > \sigma_{\bar{x}''}^2$.

This result follows from two facts:

(1) If one element is being taken from each stratum then the high average correlation between strata means results in the efficiency of the stratified random sampling mean being greater than the efficiency of the systematic sampling mean, despite the equal within stratum variances.

(2) If more than one element is being taken from each stratum then the within stratum variance of the systematic sampling mean is less than the within stratum variance of the stratified random sampling mean and if the size of stratum and sample from stratum are large enough, the smaller within stratum variance of the systematic sample more than compensates for the correlation among strata means.

Of course, in a straight line distribution there are much more efficient methods of defining a stratified random sample than that we have used. Furthermore, more efficient sampling procedures than those discussed are available. However, this example will be of use in indicating the general problems that arise as well as the procedures that may be followed in attacking them.

Another comparison of systematic and stratified random sampling may be obtained by considering the x_ν to be composed of two elements, a trend function and a periodic function so that the deviations from the trend constitute a periodic function.

Let $x_\nu = \varphi_1(\nu) + \varphi_2(\nu)$, where $\varphi_1(\nu)$ is a trend function and $\varphi_2(\nu)$ is a periodic function of period $2h$, $N = 2hQ$.

Let $\varphi_2(\nu) = y_\nu$. Then $y_j = y_{2h+j} = \cdots = y_{2h(Q-1)+j}$, $j = 1, \cdots, 2h$ and $y_{2ha+j} - \bar{y} = -(y_{2ha+h+j} - \bar{y})$, $j = 1, \cdots, h$, $a = 0, \cdots, Q - 1$.

Since the sizes of sample that we shall consider for purposes of this comparison are all multiples of h we shall calculate our variances and covariances so that we obtain all the necessary information at once.

Let the mean of $\varphi_1(\nu)$ be denoted by $\bar{\varphi}_1$ and let the mean of $\varphi_2(\nu)$ be denoted by $\bar{y}$. Then $\bar{x} = \bar{\varphi}_1 + \bar{y}$ and

$$\begin{aligned} N\sigma^2 &= \sum_\nu (x_\nu - \bar{x})^2 \\ &= \sum_\nu [\varphi_1(\nu) - \bar{\varphi}_1]^2 + \sum_\nu (y_\nu - \bar{y})^2 + 2\sum_\nu [\varphi_1(\nu) - \bar{\varphi}_1](y_\nu - \bar{y}) \\ &= \sum_{a,i} (\varphi_{1(a-1)h+i} - \bar{\varphi}_{1a})^2 + h\sum_a (\bar{\varphi}_{1a} - \bar{\varphi}_1)^2 \\ &\quad + \sum_{a,i} (y_{(a-1)h+i} - \bar{y}_a)^2 + h\sum_a (\bar{y}_a - \bar{y})^2 \\ &\quad + \sum_{a,i} [\varphi_{1(a-1)h+i} - \bar{\varphi}_{1a}](y_{(a-1)h+i} - \bar{y}_a) + h\sum_a (\bar{\varphi}_{1a} - \bar{\varphi}_1)(\bar{y}_a - \bar{y}) \end{aligned}$$

where $a = 1, \cdots, 2Q$; $i = 1, \cdots, h$, $\bar{\varphi}_{1a}$ is the arithmetic mean of $\varphi_{1[(a-1)h+1]}$, $\cdots$, $\varphi_{1[ah]}$ and $\bar{y}_a$ is the arithmetic mean of $y_{(a-1)h+1}, \cdots, y_{ah}$.

It follows from the assumptions with respect to the y_ν that $\bar{y}_a = \bar{y}$ and that $\sum_i (y_i - \bar{y}_a)^2$ is the same for each value of a and is equal to $\sum_{i=1}^{h} (y_i - \bar{y})^2$.

Since $y_i = y_{2h+i} = \cdots = y_{2h(Q-1)+i}$ and $y_{h+i} = y_{3h+i} = \cdots = y_{2h(Q-1)+h+i}$ we have

$$\sum_{\nu} [\varphi_1(\nu) - \bar{\varphi}_1](y_\nu - \bar{y}) = \sum_{i=1}^{h} (y_i - \bar{y}) \sum_{a=1}^{Q} [\varphi_{1(2a-2)h+i} - \bar{\varphi}_{12a-2}]$$

$$+ \sum_{i=h+1}^{2h} (y_i - \bar{y}) \sum_{a=1}^{Q} [\varphi_{1((2a-2)h+i)} - \bar{\varphi}_{12a-2}].$$

Since $y_i - \bar{y} = (y_{i+h} - \bar{y})$, we also have $\sum_{\nu} [\varphi_1(\nu) - \bar{\varphi}_1](y_\nu - \bar{y}) = \sum_{i=1}^{h} (y_i - \bar{y}) \left\{ \sum_{a=1}^{Q} [\varphi_1((2a - 2)h + i) - \varphi_1((2a - 1)h + i) - \bar{\varphi}_{12a-2} + \varphi_{12a-1}] \right\}$ which vanishes for example if $\varphi_1(\nu)$ is a straight line or if $\varphi_1(\nu)$ is a succession of straight lines each having length $2h$.

Let us now assume that $\varphi_1(\nu) = A + B\nu$. Then $\bar{\varphi}_{1a} = A + B\left[(a - 1)h + \frac{h + 1}{2}\right]$,

$$\varphi_1((a - 1)h + i) - \bar{\varphi}_{1a} = i - \frac{h + 1}{2},$$

$$\sum_{i} [\varphi_1((a - 1)h + i) - \bar{\varphi}_{1a}]^2 = \frac{B^2h(h^2 - 1)}{12},$$

$$\sum_{a} (\bar{\varphi}_{1a} - \bar{\varphi}_1)^2 = \frac{B^2h^2(4Q^2 - 1)(2Q)}{12},$$

$$\bar{\varphi}_1 = A + B\,\frac{2h + 1}{2},$$

$$\sum_{\nu} (\varphi_1(\nu) - \bar{\varphi}_1)^2 = \frac{2hQB^2}{12}\,[4h^2Q^2 - 1].$$

Then $\sigma^2 = \sigma_y^2 + \frac{B^2}{12}[4h^2Q^2 - 1]$ where $h\sigma_y^2 = \sum_i (y_i - \bar{y})^2$, and the variance of the mean of an unrestricted random sample of size n is $\sigma_{\bar{x}'}^2 = \frac{N - n}{N - 1}\frac{\sigma^2}{n}$.

Let us assume now that the size of stratum is mh where m is a factor of $2Q$, say $2Q/m = L_m$. Then the variance within each of the L_m strata is a constant, say, σ_1^2 where $\sigma_1^2 = \sigma_y^2 + \frac{B^2}{12}[h^2m^2 - 1]$ if m is even. If m is odd then L_m is even and in half the strata the within stratum variance is $\sigma_1^2 + \frac{1}{hm}\sum_{i=1}^{h} (y_i - \bar{y})\left(i - \frac{h + 1}{2}\right)$ while in the other strata, the within stratum variance is $\sigma_1^2 - \frac{1}{hm}\sum_{i=1}^{h} (y_i - \bar{y})\left(i - \frac{h + 1}{12}\right)$.

Then, if c elements are sampled at random from each of the L_m strata, it follows that

$$\begin{aligned}\sigma^2_{\bar{x}''} &= \frac{1}{L_m}\left(\frac{mh - c}{mh - 1}\right)\frac{\sigma_1^2}{c} \\ &= \frac{1}{L_m}\left(\frac{mh - c}{mh - 1}\right)\frac{1}{c}\left(\sigma_y^2 + \frac{B^2}{12}[h^2m^2 - 1]\right).\end{aligned}$$

In order to evaluate the variance of the systematic sampling mean let us evaluate $\sum_j (x_j - x_{j+k\delta})^2 = k(n - \delta)s'_{k\delta}$. Now upon substituting for x_ν, it follows that $k(n - \delta)s'_{k\delta} = \sum_j (y_j - y_{j+k\delta})^2 - 2Bk\delta \sum_j (y_j - y_{j+k\delta}) + k(n - \delta)B^2k^2\delta^2$.

Then, if k is a multiple of h, it follows that $\sum_j (y_j - y_{j+k\delta}) = 0$. Furthermore, if k is an even multiple of h, then $y_j = y_{j+\delta k}$ and hence $\sum_j (y_j - y_{j+\delta k})^2 = 0$. Finally, if k is an odd multiple of h then, if δ is an odd number $y_j - y_{j+\delta k} = 2(y_j - \bar{y})$ while, if δ is even $y_j - y_{j+\delta k} = 0$ and hence

$$\begin{aligned}\sum_j (y_j - y_{j+\delta k})^2 &= 4\sum_j (y_j - \bar{y})^2 \\ &= 4\frac{k(n - \delta)}{h}\sum_i (y_i - \bar{y})^2\end{aligned}$$

if k is an odd multiple of h and δ is an odd number. Note that if k is an odd multiple of h, then n is an even number. Since

$$\sigma^2_{\bar{x}} = \sigma^2 - \frac{1}{n^2}\sum_\delta (n - \delta)s'_{k\delta}$$

it is necessary to evaluate $\sum_\delta (n - \delta)s'_{k\delta}$. Now, if k is an even multiple of h, it follows that $(n - \delta)s'_{k\delta} = (n - \delta)B^2k^2\delta^2$ and

$$\begin{aligned}\sum_\delta (n - \delta)s'_{k\delta} &= B^2k^2\{n\sum_\delta \delta^2 - \sum_\delta \delta^3\} \\ &= B^2k^2\frac{n^2(n^2 - 1)}{12}\end{aligned}$$

Hence, if k is an even multiple of h, it follows that $\sigma^2_{\bar{x}} = \sigma^2 - \frac{B^2k^2(n^2 - 1)}{12}$.

On the other hand, if k is an odd multiple of h, and if δ is odd, we have $(n - \delta)s'_{k\delta} = (n - \delta)B^2k^2\delta^2 + 4(n - \delta)\sigma_y^2$ while if δ is even $(n - \delta)s'_{k\delta} = (n - \delta)B^2k^2\delta^2$.

Hence

$$\sum_\delta (n - \delta)s'_{k\delta} = \frac{B^2k^2n^2(n^2 - 1)}{12} + n^2\sigma_y^2 .$$

Hence, if k is an odd multiple of h, it follows that

$$\sigma_{\bar{x}}^2 = \sigma^2 - \frac{B^2k^2(n^2-1)}{12} - \sigma^2 .$$

Then, if k is an even multiple of h

$$\sigma_{\bar{x}}^2 = \sigma_y^2 + \frac{B^2}{12}(k^2n^2-1) - \frac{B^2}{12}k^2(n^2-1)$$

$$= \sigma_y^2 + \frac{B^2}{12}(k^2-1),$$

and, if k is an odd multiple of h, then $\sigma_{\bar{x}}^2 = \frac{B^2}{12}B^2(k^2-1)$.

Thus, systematic sampling will yield superior results if

$$c > \frac{L}{1 + \frac{12\sigma_y^2}{B^2(hm)(hm-1)} - \frac{L(L-1)}{N}} .$$

Since $\frac{N}{L} > c$ it follows that for a solution, c, to exist, we must have

$$N > 2L^2 - L - \frac{12\sigma_y^2}{B^2}\left(\frac{L^2}{N-L}\right).$$

12. Summary. In this paper we have presented the theoretical basis for systematic sampling for stratified and unstratified populations including the derivation of the variances, a study of the possible values of the parameters, estimates of the parameters, the effects of changing the size of sample, and comparisons among systematic sampling, unrestricted random sampling, and stratified random sampling. The paper contains for the case where the sampling unit consists of one element, not only the theory necessary, but in addition, some analysis of the conditions under which systematic sampling ought be used, and formulas for calculating the variances.

In later papers of this series, we shall present the theory of systematic sampling when the sampling unit is a cluster of elements, the theory when we assume we are sampling not from a finite population but an infinite population, each of whose elements is normally distributed, and further studies of various parts of the theory and practice of systematic sampling.

Appendix A

Values Assumed by Certain Variables

In order to avoid repeating the limits of summation of variables, we shall give these limits in this appendix.

TABLE I

Values Assumed by Subscripts

Letter	The letter will assume all integral values from 1 to
i	k
λ	$n - \delta$
j	$k(n - \delta)$
δ	$n - 1$
ν, ν'	kn
α	n
γ	n
μ, μ'	$n/2$ if n is even, $\frac{n-1}{2}$ if n is odd
a, b	L

The letter β will assume the values $i_1, i_2, \cdots, i_g$ where $i_1, \cdots, i_g$ are a selection of g of the k integers $1, \cdots, k$.

Appendix B

On the Limits of Some Finite Sums

The difficulties that arise in the transformation of finite sums are very similar to those that arise in the theory of transforming multiple integrals, i.e., the effects of transforming variables or order of summation on the limits of summation. Certain lemmas that have proved useful in this paper are presented separately here in a more general form.

Let $f(u)$ and $f(u, v)$ be functions of u and v that are finite for all possible values of u and v.

Lemma 1.

$$\sum_{\substack{\alpha,\gamma \\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}, x_{i+(\gamma-1)k}) = \sum_{\delta,\lambda} f(x_{i+(\lambda-1)k}, x_{i+(\lambda+\delta-1)k})$$

Proof: Let $\alpha = \lambda$ and let $\gamma = \lambda + \delta$. Since $1 \leq \alpha < \gamma$ and $\gamma \leq n$, the possible values of δ are $1, \cdots, (n - 1)$. For any fixed value of δ the possible values of λ then are 1 to $n - \delta$ since $\lambda = \gamma - \delta$ and, for a fixed value of δ the maximum value of λ is determined when $\gamma = n$. With these limits each term of f on the left side of the equation occurs once and only once on the right side of the equation. Furthermore, no additional term occurs on the right side of the equation.

Lemma 2.

$$\sum_i \sum_\lambda f(x_{i+(\lambda-1)k}, x_{i+(\lambda+\delta-1)k}) = \sum_j f(x_j, x_{j+k\delta}).$$

Proof: Let $j = i + (\lambda - 1)k$. Then j is a monotone increasing function of i and λ. The minimum value of j occurs when $i = 1$. In that case $j = 1$. The maximum value of j occurs when $i = k$, $\lambda = n - \delta$. In that case $j = (n - \delta)k$.

With these limits each term of f on the left side of the equation occurs once and only once on the right side of the equation. Furthermore, no additional term occurs on the right side of the equation.

LEMMA 3.

$$\sum_{\substack{i,\alpha,\gamma \\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}, x_{i+(\gamma-1)k}) = \sum_{\delta,j} f(x_j, x_{j+\delta k}).$$

PROOF: First apply Lemma 1 to $\sum_{\substack{\alpha,\gamma \\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}, x_{i+(\gamma-1)k})$ and then apply Lemma 2 to the resulting expression.

LEMMA 4.

$$\sum_{\delta,j} [f(x_j) + f(x_{j+k\delta})] = (n-1) \sum_{\nu} f(x_\nu).$$

PROOF: Let $m = j + k\delta$. Then for any fixed value of δ the minimum value of m occurs when $j = 1$. In that case $m = k\delta + 1$. For any fixed value of δ, the maximum value of m occurs when $j = k(n - \delta)$. In that case $m = nk$. The letter m will assume all integral values from $k\delta + 1$ to kn, and hence,

$$\sum_{\delta,j} f(x_j) + \sum_{\delta,j} f(x_{j+k\delta}) = \sum_{\delta,j} f(x_j) + \sum_{\delta,m} f(x_m).$$

If we sum δ from $n - 1$ to 1 instead of from 1 to $n - 1$ in $\sum_{\delta,m} f(x_m)$ we see that

$$\begin{aligned} \sum_{\delta,j} f(x_j) + \sum_{\delta,m} f(x_m) = &\sum_{j=1}^{k(n-1)} f(x_j) + \sum_{m=k(n-1)+1}^{kn} f(x_m) \\ &+ \cdots \\ &+ \sum_{j=1}^{k} f(x_j) + \sum_{m=k+1}^{kn} f(x_m) \end{aligned}$$

where the summations of x_j are terms of $\sum_{\delta,j} f(x_j)$ and the summations of x_m are terms of $\sum_{\delta,m} f(x_m)$. But $\sum_{j=1}^{k(n-\delta)} f(x_j) + \sum_{m=k(n-\delta)+1}^{kn} f(x_m) = \sum_{\nu} f(x_\nu)$ and hence Lemma 4 is proved.

LEMMA 5. *Let*

$$\sum_{i,\alpha,\gamma} f(x_{i+(\alpha-1)k}) f(x_{i+(\gamma-1)k}) = A.$$

Then

$$A = n \sum_{\nu} [f(x_\nu)]^2 - \sum_{j,\delta} [f(x_j) - f(x_{j+k\delta})]^2.$$

PROOF:

$$A = \sum_{i,\alpha} [f(x_{i+(\alpha-1)k})]^2 + 2 \sum_{\substack{i,\alpha,\gamma \\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}) f(x_{i+(\gamma-1)k}).$$

By Lemma 3

$$2 \sum_{\substack{i,\alpha,\gamma \\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}) f(x_{i+(\gamma-1)k}) = 2 \sum_{\delta,j} f(x_j) f(x_{j+\delta k})$$

and since we have

$$2f(x_j)f(x_{j+\delta k}) = f(x_j)^2 + f(x_{j+\delta k})^2 - [f(x_j) - f(x_{j+\delta k})]^2,$$

the proof is completed by using Lemma 4.

LEMMA 6. Let $kn\bar{f} = \sum_\nu f(x_\nu)$. *Then*

$$A\left(\frac{1}{kn^2}\right) - \bar{f}^2 = \left(\frac{1}{kn}\right) \sum_\nu [f(x_\nu) - \bar{f}] - \left(\frac{1}{kn^2}\right) \sum_{j,\delta} [f(x_j) - f(x_{j+k\delta})|^2.$$

PROOF: This lemma is a direct consequence of Lemma 5.

LEMMA 7.

$$A = \sum_\nu [f(x_\nu) - \bar{f}]^2 + 2\sum_{j,\delta} [f(x_j) - \bar{f}][f(x_{j+k\delta}) - \bar{f}] + kn^2\bar{f}^2.$$

PROOF: From Lemma 4, it follows that

$$A = n\sum_\nu f(x_\nu)^2 - \sum_{j,\delta} \{[f(x_j) - \bar{f}]^2 + [f(x_{j+k\delta}) - \bar{f}]^2 + 2\sum_{j,\delta} [f(x_j) - \bar{f}][f(x_{j+k\delta}) - \bar{f}]$$

and hence, from Lemma 3, it follows that

$$A = n\sum_\nu [f(x_\nu) - \bar{f}]^2 + n^2k\bar{f}^2 - (n-1)\sum_\nu [f(x_\nu) - \bar{f}]^2$$
$$+ 2\sum_{j,\delta}[f(x_j) - \bar{f}][f(x_{j+k\delta}) - \bar{f}],$$

whence the lemma is proved.

LEMMA 8.

$$A\left(\frac{1}{kn^2}\right) - \bar{f}^2 = \left(\frac{1}{kn^2}\right) \sum_\nu [f(x_\nu) - \bar{f}]^2 + \left(\frac{2}{kn^2}\right) \sum_{j,\delta} [f(x_j) - \bar{f}]\,[f(x_{j+k\delta}) - \bar{f}].$$

This lemma is a direct consequence of Lemma 7.

LEMMA 9. *If* $h > kn$, *let* x_h *equal* x_{h-kn}. *Let* $f(u, v) = f(v, u)$ *i.e.* f *is symmetric. Then, if we let*

$$d_{k\delta} = \sum_j f(x_j, x_{j+k\delta})$$

it follows that

$$d_{k\delta} + d_{kn-\delta} = \sum_\nu f(x_\nu, x_{\nu+k\delta}).$$

PROOF: Obviously

$$\sum_\nu f(x_\nu, x_{\nu+k\delta}) = d_{k\delta} + B,$$

where

$$B = \sum_{g=k(n-\delta)+1}^{kn} f(x_g, x_{g+k\delta}).$$

Now, let $h = g - (n - \delta)k$. Then

$$B = \sum_{h=1}^{\delta k} f(x_{h+(n-\delta)k}, x_{h+kn}).$$

Since $x_{h+kn} = x_h$, and $f(x_{h+(n-\delta)k}, x_h) = f(x_h, x_{h+(n-\delta)k})$, it follows that $B = d_{kn-\delta}$ and the lemma is proved. It is noted that the symmetry of $f(u, v)$ is necessary as well as sufficient, for if $f(x_\nu, x_{\nu+k\delta}) = x_\nu - x_{\nu+k\delta}$ the theorem is false.

Appendix C

Stratified Sampling

Let the population P consist of L strata $P_1, \cdots, P_L$. Let $\bar{x}$ be the arithmetic mean of P, and $\bar{x}_a$ the arithmetic mean of P_a. Let $\tilde{x}_a$ be the sample estimate of $\bar{x}_a$, and let $\tilde{x} = \sum_a c_a\tilde{x}_a$. Then $\mathcal{E}\tilde{x} = \sum_a c_a A_a = A$ where $\mathcal{E}\tilde{x}_a = A_a$. Let $\sigma_{\tilde{x}}^2$ be defined by $\sigma_{\tilde{x}}^2 = \mathcal{E}(\tilde{x} - \bar{x})^2$. Then $\sigma_{\tilde{x}}^2 = \mathcal{E}(\tilde{x} - A)^2 + (A - \bar{x})^2$ and hence it is easy to see that $\sigma_{\tilde{x}}^2 = \sum_{a,b} c_a c_b \sigma_{\tilde{x}_a\tilde{x}_b} + (A - \bar{x})^2$ where

$$\sigma_{\tilde{x}_a\tilde{x}_b} = \mathcal{E}(\tilde{x}_a - A_a)(\tilde{x}_b - A_b),$$

$$(A - \bar{x})^2 = \left[\sum_a \left(c_a A_a - \frac{N_a}{N}\bar{x}_a\right)\right]^2,$$

and if $NC_a = N_a$, then

$$(A - \bar{x})^2 = \sum_{a,b} c_a c_b (A_a - \bar{x}_a)(A_b - \bar{x}_b)$$

and $\sigma_{\tilde{x}}^2 = \sum_{a,b} c_a c_b \sigma_{\tilde{x}_a\tilde{x}_b}$

where

$$\sigma_{\tilde{x}_a\tilde{x}_b}^2 = \mathcal{E}(\tilde{x}_a - \bar{x}_a)(\tilde{x}_b - \bar{x}_b).$$

These formulae hold whatever may be the method used in sampling the ith stratum. If $\tilde{x}$ is an unbiased estimate of $\bar{x}$ and $\tilde{x}_a$ is independent of $\tilde{x}_b$, then the usual formula $\sigma_{\tilde{x}}^2 = \sum_a c_a^2\sigma_{\tilde{x}_a}^2$ holds. The formula for $\sigma_{\tilde{x}_a\tilde{x}_b}$ will, of course, depend on whether a random, cluster, systematic, or other sampling procedure is used.