TWO DIMENSIONAL SYSTEMATIC SAMPLING AND THE ASSOCIATED STRATIFIED AND RANDOM SAMPLING

By A. C. DAS

Statistical Laboratory, Calcutta

INTRODUCTION

Systematic sampling has now come to be of much interest both in theoretical and applied statistics. This type of sampling with its various modifications is well known. Madow (1944), Cochran (1946) and Yates (1948) have discussed some particular types in one dimension. This paper deals with the two dimensional generalisation of Cochran's results for the uni-dimensional case.

Let the universe to be sampled consist of ml rows, each of nk cells in mn strata, each stratum containing lk cells. The cells are square or rectangular as the case may be. Let the co-ordinates (x,y) be taken in such a manner that, $y =$ constant, gives a row and $x =$ constant, gives a column. Thus in each row n sets of k cells are along x direction and in each column m sets of l cells are along y direction. In a particular stratum, ν cells out of lk cells are chosen at random. Then every lth row of cells (along y direction), beginning from the one in which each of these chosen cells occurs, is selected. On each of these selected rows, cells in same column with the randomly chosen ones and every kth cell (along x direction) are taken, thus comprising a sample of size $mn\nu$, which is a kind of systematic sampling in two dimensions. This type of systematic sampling may also be defined as: ν cells out of lk cells are taken at random from a particular stratum; then in every other stratum a similar set of ν cells are chosen, having the same configuration within the set and also with respect to the stratum as in the first stratum. We may also take a random sample of size ν from each of the strata forming a total stratified sample of size $mn\nu$. Alternatively, $mn\nu$ cells may be taken at random out of all $mnlk$ cells without stratification, which will constitute a perfectly random sample.

We assume this universe to be homogeneous in the sense that $E(z_{ij}) = \mu$, where z's are the observables of the cells, and that observables have got common variance and different space-correlations. Let $\rho(u,v)$ be the space-correlation between points separated by distances u and v along x and y directions respectively, *i.e.*, if P be the cell (x,y), we shall pair its yield against the yield of the cell $(x+u, y+v)$ and from the set of all such possible pairs shall calculate $\rho(u, v)$, which may have certain fluctuations due to the finiteness of the universe. If this universe be infinite, such an effect will be absent. If, however, the universe be finite it may be considered to be a sample from an infinite population, as has been done by Cochran (1946), where $E(z) = \mu$, $V(z) = \sigma^2$, and $E\{(z_{i+u,j+v} - \mu)(z_{ij} - \mu)\} = \rho(u,v)$ for the infinite population. This $\rho(u,v)$ for different values of u and v will form a solid figure which may be called a correlopiped. This figure will be a continuous one for infinitesimal sizes of the cells.

In this paper the expected values of the variances of the grand sample mean, as averaged over all finite populations from the infinite population, under random, stratified and systematic sampling will be denoted by σ_r^2, σ_{st}^2 and σ_{sy}^2 respectively following the symbols used by Cochran (1946). $\Delta_1, \Delta_2, \delta_1, \delta_2$ are the operators, such that

$$\Delta_1\rho(u,v) = \rho(u+1,v) - \rho(u,v),$$
$$\Delta_2\rho(u,v) = \rho(u,v+1) - \rho(u,v),\ \delta_1\rho(u,v) = \rho(u+\tfrac{1}{2},v) - \rho(u-\tfrac{1}{2},v),$$
$$\delta_1^2\rho(u,v) = \rho(u+1,v) + \rho(u-1,v) - 2\rho(u,v),\ \Delta_1\Delta_2\rho(u,v) = \Delta_1\rho(u,v+1) - \Delta_1\rho(u,v)$$

and similar definitions for δ_2, δ_2^2 and $\delta_1^2\,\delta_2^2$. We shall also use E_1, E_2 such that $E_1\rho(u,v) = \rho(u+1,v)$ and $E_2\rho(u,v) = \rho(u,v+1)$. These E_i's are different from E (with no subscript) which denotes "expectation". It is also to be noted that $\rho(u,v) = \rho(-u,-v)$ and $\rho(u,-v) = \rho(-u,v)$ are different and we shall denote $\rho(u,v)+\rho(-u,+v)$ by $\psi(u,v)$ which is symmetrical with respect to u and v. $\Delta_1\psi$ and $\Delta_2\psi$ will always denote $\Delta_1\psi(u,v)$ and $\Delta_2\psi(u,v)$ respectively.

In section 1, σ_r^2, σ_{st}^2 and σ_{sy}^2 have been calculated and it has been shown that their relative proportions are independent of ν, and hence the relative efficiencies of the three types of sampling. In section 2 a set of sufficient conditions under which stratified sampling is more efficient than random sampling has been stated in a number of theorems, and in section 3, a sufficient condition under which systematic sampling is more efficient than stratified sampling has been stated in Theorem 4. From the results of these two sections, we can find out the set of sufficient conditions under which systematic sampling is more efficient than random sampling; though the two have not been directly compared. Cochran's results come out as special cases. But there are some improvements over Cochran (1946) that the results are true whatever may be ρ, positive or negative, and that ν may be greater than unity.

Section 1

It can be proved that $\sum\limits_{i=1}^{n}(z_i-\bar{z})^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{i<j=1}^{n}(z_i-z_j)^2$ ... (1.1)

Lemma 1. In a plot containing M rows of N cells each, with z_{ij} as the stochastic variate corresponding to the (i,j)th cell, $i = 1,2, \ldots, N$; $j = 1,2, \ldots, M$.

$$E\left\{\sum_{i=1}^{N}\sum_{j=1}^{M}(z_{ij}-\bar{z})^2\right\} = \frac{2\sigma^2}{MN}\left[\frac{MN(MN-1)}{2} - N\sum_{v=1}^{M-1}(M-v)\rho(0,v) - M\sum_{u=1}^{N-1}(N-u)\rho(u,0)\right.$$
$$\left. - \sum_{u=1}^{N-1}\sum_{v=1}^{M-1}(N-u)(M-v)\psi(u,v)\right]$$

Proof. By (1.1), $\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M}(z_{ij}-\bar{z})^2 = \frac{1}{MN}\Sigma(z_t-z_{t'})^2$ where the summation is to be extended over all combinations of z_t and $z_{t'}$ in the finite plot, which is a sample from an infinite plot of the type mentioned in the introduction. There are ${}^{MN}c_2$ combinations in all splitted as: (i) within each row ${}^{N}c_2$ combinations, being $M.\ {}^{N}c_2$ in all, (ii) within each column ${}^{M}c_2$, in all $N\,{}^{M}c_2$ and (iii) between (x, y) and $(x+u, y+v)$, $(u \neq 0, v \neq 0)$, giving $2\,{}^{M}c_2\,{}^{N}c_2$.

Now, $$E\left\{\sum_{i,j}(z_{ij}-\bar{z})^2\right\} = \frac{1}{MN}E\{\Sigma(z_t-\mu+\mu-z_{t'})^2\} = \frac{2\sigma^2}{MN}\Sigma\{1-\rho(z_t, z_{t'})\} \qquad \dots \quad (1.2)$$

where $\rho(z_t, z_{t'})$ means the correlation between z_t and $z_{t'}$ and the summation is to be carried over all combinations of t and t' (*i.e.*, ${}^{MN}c_2$ in all). Let us first find $\Sigma\rho(z_t, z_{t'})$. It will be the sum of three expressions corresponding to the three cases (i), (ii) and (iii) mentioned above. In case (i) there are $(N-u)$ cells $(u,0)$–distance apart in each row; therefore, $\Sigma\rho(z_t, z_{t'})$ of this type over all rows is $M\sum\limits_{u=1}^{N-1}(N-u)\rho(u,0)$. Similarly for case (ii), $\Sigma\rho(z_t, z_{t'})$ is $N\sum\limits_{v=1}^{M-1}(M-v)\rho(0, v)$. In case (iii) considering the four points (x, y), $(x+u, y)$, $(x, y+v)$ and $(x+u, y+v)$ with $u \neq 0$ and $v \neq 0$, it is obvious that there will be only 2 effective types of distances of case (iii) among these four points, namely, (u,v) and $(-u, v)$. The distance $(-u, -v)$ is equivalent to (u, v), since the space correlation will be the same for both the distances ; similarly, the distance $(u, -v)$ is equivalent to $(-u, v)$. Therefore, summing $\Sigma\rho(z_t, z_{t'})$ over this type we get $\sum\limits_{u=1}^{N-1}\sum\limits_{v=1}^{M-1}(N-u)(M-v)\psi(u,v)$; beacause there are $(N-u)\,(M-v)$ sets of such quadruplets. It is a point to note that $\rho(u, 0) = \rho(-u, 0)$ and $\rho(0, v) = \rho(0,-v)$.

Again $\Sigma 1$ over all combinations is ${}^{MN}c_2$.

Then the lemma follows from (1.2), which can be written as:

$$E\left\{\sum_{i=1}^{N}\sum_{j=1}^{M}(z_{ij}-\bar{z})^2\right\} = (MN-1)\sigma^2\{1-\phi(M, N)\} = (MN-1)\sigma^2 L(M, N) \qquad \dots \quad (1.3)$$

where $$\phi(M, N) = \frac{2}{M(MN-1)}\sum_{v=1}^{M-1}(M-v)\rho(0, v)+\frac{2}{N(MN-1)}\sum_{u=1}^{N-1}\rho(u, 0)$$
$$+\frac{2}{MN(MN-1)}\sum_{u=1}^{N-1}\sum_{v=1}^{M-1}(N-u)\,(M-v)\psi(u, v) \qquad \dots \quad (1.4)$$

and $$L(M, N) = 1-\phi(M, N) \qquad \dots \quad (1.5)$$

$\therefore$ putting $M = ml$ and $N = nk$ in (1.3), we have for the whole finite population:

$$E\{\Sigma(z-\bar{z})^2\} = (mnlk-1)\sigma^2\{1-\phi(ml, nk)\} \qquad \dots \quad (1.6)$$

Similarly, for each stratum, putting $M = l$ and $N = k$ in (1.3)

$$E\{\Sigma(z-\bar{z})^2\} = (lk-1)\sigma^2\{1-\phi(l, k)\} \qquad \dots \quad (1.7)$$

Therefore, for mn strata, E (s.s. within strata) $= mn(lk-1)\sigma^2\{1-\phi(l, k)\}$ $\quad \dots \quad (1.8)$

For each of the systematic samples for $\nu=1$,
$E\{\Sigma(z-\bar{z})^2\} = (mn-1)\sigma^2\{1-\phi_{sy}(m,n)\}$ putting $M=m$ and $N=n$, replacing $\rho(u,v)$ by $\rho(uk, vl)$ in (1.3) and (1.4) and replacing ϕ by ϕ_{sy}.

$$\therefore \text{ for } \nu=1,\ E\,(\text{s.s. within systematic sample}) = lk(mn-1)\sigma^2\{1-\phi_{sy}(m,n)\} \quad \ldots \quad (1.9)$$

Now, for a random sample size n' from a finite population of size N, the variance of sample mean $= \dfrac{1}{n'}\,\dfrac{N-n'}{N-1}\left\{\dfrac{1}{N}\,\Sigma(z-\bar{z})^2\right\}$, summation being taken over whole finite population.

Hence putting $N = mnlk$ in our case, $n' = mn\nu$, we have, using (1.6) the expected variance of sample mean as:

$$\sigma_r^2 = \frac{1}{mn\nu}\,\frac{mn(lk-\nu)}{(mnlk-1)}\,\frac{1}{mnlk}\,(mnlk-1)\sigma^2\{1-\phi(ml,\ nk)\} = \frac{\sigma^2}{mn\nu}\left(1-\frac{\nu}{lk}\right)\left\{1-\phi(ml,\ nk)\right\} \quad \ldots \quad (1.10)$$

Again, there are mn strata with lk cells in each and let $x_{i1}, x_{i2}, \ldots, x_{i\nu}$ be the sample of size ν from the ith stratum. Here x_{ij} is a stochastic variate and is not to be confused with x, having no subscripts, that denotes a co-ordinate in the introduction. Then

$$\bar{x} = \frac{1}{mn\nu}\sum_{i=1}^{mn}\sum_{j=1}^{\nu} x_{ij} = \frac{1}{mn}\sum_{i=1}^{mn}\bar{x}_i\,,\quad \text{where } \bar{x}_i = \frac{1}{\nu}\sum_{j=1}^{\nu} x_{ij}\ ;\quad V(\bar{x}) = \frac{1}{m^2n^2}\sum_{i=1}^{mn} V(\bar{x}_i).$$

But $E\{V(\bar{x}_i)\} = \dfrac{1}{\nu}\,\dfrac{(lk-\nu)}{(lk-1)}\,\dfrac{(lk-1)}{lk}\,\sigma^2\{1-\phi(l,\ k)\}$, from (1.7).

Hence
$$\sigma_{st}^2 = E\{V(\bar{x})\} = \frac{\sigma^2}{mn\nu}\left(1-\frac{\nu}{lk}\right)\left\{1-\phi(l,\ k)\right\} \quad \ldots \quad (1.11)$$

Now, to find the expected variance σ_{sy}^2 of systematic sample, let $x_{11}, x_{12}, \ldots, x_{1\nu}$ be the sample from the first stratum. Then the samples from other strata are automatically fixed up. The grand sample mean $\bar{x}_{sy} = \dfrac{1}{mn\nu}\displaystyle\sum_{i=1}^{mn}\sum_{j=1}^{\nu} x_{ij}$ and $\sigma_{sy}^2 = E(\bar{x}_{sy}-\bar{\bar{x}})^2$, where $\bar{\bar{x}}$ is the grand mean of the finite population.

Here let $\displaystyle\sum_{i=1}^{mn} x_{ij}/mn$ be denoted by ${}_j\bar{x}_{sy}$, so that $(\bar{x}_{sy}-\bar{\bar{x}}) = \dfrac{1}{\nu}\displaystyle\sum_{j=1}^{\nu}({}_j\bar{x}_{sy}-\bar{\bar{x}})$

Hence
$$(\bar{x}_{sy}-\bar{\bar{x}})^2 = \frac{1}{\nu^2}\cdot\left\{\sum_{j=1}^{\nu}({}_j\bar{x}_{sy}-\bar{\bar{x}})^2 + \sum_{\substack{j,j'=1\\(j\neq j')}}^{\nu}({}_j\bar{x}_{sy}-\bar{\bar{x}})\,({}_{j'}\bar{x}_{sy}-\bar{\bar{x}})\right\} \quad \ldots \quad (1.12)$$

We shall take the average of (1.12) over ${}^{lk}c_\nu$ systematic samples.

If x_{ij} be a random point (observable) from the first stratum, ${}_j\bar{x}_{sy}$ is the mean of the corresponding systematic sample. Then for summation of (1.12) over ${}^{lk}c_\nu$ systematic samples, any ${}_j\bar{x}_{sy}$ $(j = 1, 2, \ldots, mn)$, will occur ${}^{lk-1}c_{\nu-1}$ times, because only $(\nu-1)$ will be required from the remaining $(lk-1)$ to complete a sample of size ν; and ${}_j\bar{x}_{sy}$ and ${}_{j'}\bar{x}_{sy}$, $(j \neq j')$, will occur together ${}^{lk-2}c_{\nu-2}$ times; therefore in the expression (1.12), when summed over ${}^{lk}c_\nu$ samples, $({}_j\bar{x}_{sy}-\bar{x}), j = 1, 2, \ldots, lk$, will occur ${}^{lk-1}c_{\nu-1}$ times and

$$({}_j\bar{x}_{sy}-\bar{x})({}_{j'}\bar{x}_{sy}-\bar{x})(j \neq j'),\ {}^{lk-2}c_{\nu-2} \text{ times. But } \sum_{i=1}^{lk}({}_i\bar{x}_{sy}-\bar{x}) = 0,$$

$$\text{whence } \sum_{i=1}^{lk}({}_j\bar{x}_{sy}-\bar{x})({}_i\bar{x}_{sy}-\bar{x}) = 0\text{; and } \sum_{j=1}^{lk}\sum_{\substack{i=1\\ i\neq j}}^{lk}({}_j\bar{x}_{sy}-\bar{x})({}_i\bar{x}_{sy}-\bar{x}) = -\sum_{j=1}^{lk}({}_j\bar{x}_{sy}-\bar{x})^2$$

Hence in the summation, $({}_j\bar{x}_{sy}-\bar{x})^2$ occurs $({}^{lk-1}c_{\nu-1} - {}^{lk-2}c_{\nu-2}) = {}^{lk-2}c_{\nu-1}$ times; and the required average of (1.12) is $\dfrac{{}^{lk-2}c_{\nu-1}}{{}^{lk}c_\nu}\,\dfrac{1}{\nu^2}\displaystyle\sum_{j=1}^{lk}({}_j\bar{x}_{sy}-\bar{x})^2 = \dfrac{1}{mn\nu}\,\dfrac{(lk-\nu)}{lk(lk-1)}\displaystyle\sum_{j=1}^{lk} mn({}_j\bar{x}_{sy}-\bar{x})^2$

$$= \frac{1}{mn\nu}\,\frac{(lk-\nu)}{lk(lk-1)}\left\{\sum_{i=1}^{mn}\sum_{j=1}^{lk}(x_{ij}-\bar{x})^2-(\text{s.s. within systematic sample with } \nu = 1)\right\} \quad \ldots(1.13)$$

Hence taking expectation

$$\sigma_{sy}^2 = \frac{\sigma^2}{mn\nu}\,\frac{(lk-\nu)}{lk(lk-1)}\left\{(mnlk-1)-(mnlk-1)\phi(ml, nk)-lk(mn-1)+lk(mn-1)\phi_{sy}(m,n)\right\}$$

from (1.6) and (1.9)

$$= \frac{\sigma^2}{mn\nu}\left(1-\frac{\nu}{lk}\right)\left\{1-\frac{(mnlk-1)}{(lk-1)}\,\phi(ml,nk) + \frac{lk(mn-1)}{(lk-1)}\,\phi_{sy}(m,n)\right\} \quad \ldots \quad (1.14)$$

and from (1.10) and (1.11):

$$\sigma_r^2 = \frac{\sigma^2}{mn\nu}\left(1-\frac{\nu}{lk}\right)\left\{1-\phi(ml,nk)\right\} \quad \ldots \quad (1.15)$$

$$\sigma_{st}^2 = \frac{\sigma^2}{mn\nu}\left(1-\frac{\nu}{lk}\right)\left\{1-\phi(l,k)\right\} \quad \ldots \quad (1.16)$$

where $\phi\,(M, N)$ is given in (1.4), and

$$\phi_{sy}(m, n) = \frac{2\,.}{m(mn-1)}\sum_{v=1}^{m-1}(m-v)\rho(0, vl) + \frac{2}{n(mn-1)}\sum_{u=1}^{n-1}\rho(uk, 0)$$

$$+ \frac{2}{mn(mn-1)}\sum_{u=1}^{n-1}\sum_{v=1}^{m-1}(n-u)(m-v)\rho(uk, vl)$$

It is thus seen that the relative efficiencies $\frac{1}{\sigma_r^2} : \frac{1}{\sigma_{st}^2} : \frac{1}{\sigma_{sy}^2}$ are independent of v and are dependent on the nature of stratification and the correlopiped, of which, if the latter be known, the former can be adjusted to suit the purpose.

Section 2

Here we shall find out a set of sufficient conditions under which stratified sampling is more efficient than random sampling, *i.e.*, under which $\sigma_r^2 \geqslant \sigma_{st}^2$. We shall at first prove a few important lemmas.

Lemma II. For all values of l and k, $L(l+1, k) \geqslant L(l, k)$, if the following conditions be satisfied ; (i) $\Delta_1\psi(u, v) < 0$, (ii) $\Delta_2\psi(u, v) < 0$, and (iii) $2\psi(u+1, v) \geqslant \psi(u, v+1) + \psi(u+1, v+1)$.

Proof. Replaceing M by l and N by k in (1.4) and (1.5), we get $L(l,k)$.
Then, putting $l' = l+1$, $A = lk-1$, $A' = l'k-1$, and

$$\beta_{0v} = 2\left(\frac{l-v}{lA} - \frac{l'-v}{l'A'}\right), \alpha_{u0} = \frac{k-u}{AA'}, \alpha_{uv} = \frac{(k-u)}{k}\, 2\left(\frac{l-v}{lA} - \frac{l'-v}{l'A'}\right) \quad \dots \quad (2.1)$$

where $u \neq 0$ and $v \neq 0$ we have

$$L(l+1; k) - L(l,k) = \sum_{v=1}^{l} \beta_{0v}\rho(0,v) + \sum_{u=1}^{k}\sum_{v=0}^{l} \alpha_{uv}\psi(u,v) \quad \dots \quad (2.2)$$

Now because $\psi(u,v) = -\Delta_2\psi(u,v) + \psi(u,v+1)$, therefore

$$\sum_{v=0}^{l} \alpha_{uv}\,\psi(u,v) = -\sum_{v=0}^{l-1} S_{uv}\Delta_2\psi(u,v) + S_{ul}\psi(u,l) \quad \dots \quad (2.3)$$

where

$$S_{uv} = \sum_{j=0}^{v} \alpha_{uj}$$

And

$$\sum_{v=1}^{l} \beta_{0v}\rho(0,v) = -\sum_{v=1}^{l-1} S'_{0v}\Delta_2\rho(0,v) + S'_{0l}\rho(0,l) \quad \dots \quad (2.4)$$

where

$$S'_{0v} = \sum_{j=1}^{v} \beta_{0j}$$

Then $S'_{0v} = v\left\{ y\left(\frac{1}{lA} - \frac{1}{l'A'}\right) + \frac{(k+1)}{l'AA'}\right\}$, where $y = l-1-v$.

Hence for

$$y \geqslant 0,\ i.e.,\ v \leqslant (l-1),\ S'_{0v} \geqslant 0 \quad \dots \quad (2.5)$$

and

$$S'_{0l} = -\frac{(k-1)}{AA'} \quad \dots \quad (2.6)$$

Now $S_{uv} = \alpha_{u0} + \sum_{j=1}^{v} \alpha_{uj} = \frac{k-u}{AA'} + \frac{k-u}{k} S'_{0v}$, which from (2.5), is positive for $v \leqslant (l-1)$,

and using (2.6),

$$S_{ul} = \frac{(k-u)}{KAA'} \quad \dots \quad (2.7)$$

Then from (2.3), (2.4), (2.5), (2.6) and (2.7), the equation (2.2) leads to

$$L(l+1,k)-L(l,k) = -\sum_{u=1}^{k}\sum_{v=0}^{l-1}\frac{(k-u)}{AA'}\,\Delta_1\psi(u,v)-\sum_{u=1}^{k}\sum_{v=0}^{l-1}\frac{(k-u)}{k}\,S'_{0v}\Delta_2\psi(u,v)$$

$$-\sum_{v=1}^{l-1}S'_{0v}\Delta_2\rho(0,v)+\sum_{u=1}^{k}\frac{(k-u)}{kAA'}\psi(u,l)-\frac{(k-1)}{2AA'}\psi(0,l) \quad \ldots \quad (2.8)$$

Now $\sum_{u=1}^{k}\frac{(k-u)}{kAA'}\,\psi(u,l)-\frac{(k-1)}{2AA'}\psi(0,l)=\sum_{u=1}^{k}\frac{(k-u)}{kAA'}\left\{\psi(u,l)-\psi(0,l)\right\}=\sum_{u=1}^{k}\frac{(k-u)}{kAA'}\sum_{i=0}^{u-1}\Delta_1\psi(i,l)$ … (2.9)

$$=\sum_{u=0}^{k}\frac{(1-u)(k-u)}{2kAA'}\,\Delta_1\psi(u-1,l)+\sum_{u=1}^{k}\frac{(k-u)}{2AA'}\,\Delta_1\psi(u-1,l) \quad \ldots \quad (2.10)$$

For $\Delta_1\psi \leqslant 0$, the first part of (2.10) is greater than or equal to zero, but the other part is not positive ; therefore for $\Delta_2\psi \leqslant 0$, and $\Delta_1\psi \leqslant 0$, we have from (2.8):

$$L(l+1,k)-L(l,k) = Q+\sum_{u=1}^{k}\frac{(k-u)}{2AA'}\,\Delta_1\psi(u-1,l)-\sum_{u=1}^{k}\sum_{v=0}^{l-1}\frac{(k-u)}{AA'}\,\Delta_2\psi(u,v) \quad \ldots \quad (2.11)$$

where $Q \geqslant 0$.

Again $\sum_{v=0}^{l-1}\Delta_2\psi(u,v) = \psi(u,l)-\psi(u,0)$.

Hence, from (2.11), $L(l+1,k) - L(l,k) = Q+\sum_{u=1}^{k}\frac{(k-u)}{2AA'}\{2\psi(u,0)-\psi(u,l)-\psi(u-1,l)\}$ … (2.12)

Now, $2\psi(u,0)-\psi(u,l)-\psi(u-1,l) \geqslant 2\psi(u,l-1)-\psi(u,l)-\psi(u-1,l)$, for $\Delta_2\psi \leqslant 0$.

$\therefore$ if $2\psi(u+1,v) \geqslant \psi(u,v+1)+\psi(u+1,v+1)$, for all values of v,

$2\psi(u,l-1)-\psi(u-1,l)-\psi(u,l) \geqslant 0$, *i.e.*, expression (2.12) and hence (2.11) is greater than or equal to zero.

Hence, the lemma is proved.

In a similar way, the following lemma can be proved.

Lemma III. For all values of l and k, $L(l,k+1) \geqslant L(l,k)$, if the following conditions be satisfied: (i) $\Delta_1\psi(u,v) \leqslant 0$, (ii) $\Delta_2\psi(u,v) \leqslant 0$, and (iii) $2\psi(u+1,v) \geqslant \psi(u+1,v)+\psi(u+1,v+1)$.

Theorem 1. For all infinite populations in which (i) $\Delta_1\psi \leqslant 0$, (ii) $\Delta_2\psi \leqslant 0$, (iii) $2\psi(u,v+1) \geqslant \psi(u+1,v)+\psi(u+1,v+1)$ and (iv) $2\psi(u+1,v) \geqslant \psi(u,v+1)+\psi(u+1,v+1)$, $\sigma_{st}^2 \leqslant \sigma_r^2$ for any size of the sample ; and $\sigma_{st}^2 < \sigma_r^2$ unless equality holds in each of the above four cases.

Proof. These four conditions satisfy all the conditions of Lemma II and Lemma III;

$$\therefore \qquad L(l, k) \leqslant L(l+1, k) \leqslant L(l+1, k+1) \leqslant \ldots \leqslant L(ml, nk)$$

$$i.e., \qquad 1-\phi(l, k) \leqslant 1-\phi(ml, nk), \text{ from (1.5).}$$

Hence, $\sigma_{st}^2 \leqslant \sigma_r^2$, from (1.10) and (1.11).

Thus the theorem is proved.

Again, if $\Delta_1\psi = \Delta_2\psi$, the conditions (iii) and (iv) of the above theorem are satisfied. Hence, a set of more stringent conditions can be stated as:

Corollary. For all infinite populations in which $\Delta_1\psi = \Delta_2\psi \leqslant 0$, $\sigma_{st}^2 \leqslant \sigma_r^2$ for any size of the sample, and $\sigma_{st}^2 < \sigma_r^2$ unless equality holds in each of the above cases.

Theorem 2. For all infinite populations in which stratification is made by parallel strips along u-direction (say), $\sigma_r^2 \geqslant \sigma_{st}^2$ for any size of the sample, if the following conditions be satisfied : (i) $\Delta_1\psi \leqslant 0$, (ii) $\Delta_2\psi \leqslant 0$, and (iii) $2\psi(u+1, v) \geqslant \psi(u, v+1) + \psi(u+1, v+1)$; and $\sigma_r^2 > \sigma_{st}^2$ unless equality holds in each of the above three cases.

Proof. These three conditions satisfy all the conditions of Lemma II.

Then $\qquad L(l, k) \leqslant L(l+1, k) \leqslant \ldots \leqslant L(ml, k)$, *i.e.*, $1-\phi(l, k) \leqslant 1-\phi(ml, k)$

Here, of course, k represents the total number of cells in a stratum; therefore

$\sigma_{st}^2 \leqslant \sigma_r^2$ and the theorem follows.

Now, condition (iii) is satisfied, if $\Delta_1\psi \geqslant \Delta_2\psi$. Hence a set of more stringent conditions can be stated as:

Corrolary 1. For all infinite populations in which stratification is made by parallel strips along u-direction (say), $\sigma_r^2 \leqslant \sigma_{st}^2$ if $\Delta_2\psi \leqslant \Delta_1\psi \leqslant 0$; and $\sigma_r^2 > \sigma_{st}^2$ unless equality holds in each of these cases.

Corrolary 2. For $k = 1$ in the above case, *i.e.*, for the one-dimensional field along v-direction, $\sigma_r^2 \geqslant \sigma_{st}^2$, if $\Delta_2\psi(u.v) \leqslant 0$, or, $\Delta_2\rho(0,v) \leqslant 0$; because here $u = 0$. This case has been discussed by Cochran (1946) for $\nu = 1$. All these theorems give the conditions irrespective of the values of l,m,n,k and hence any pattern of stratification is more efficient under them. But the author (1949) has given another set of conditions that depends on some relations of the parameters. We shall now discuss it. Here also, we have got some degrees of freedom in stratification. It can be stated in the following theorem.

Theorem 3. For all patterns of stratification in which $\sqrt{l} \leqslant k \leqslant l^2$ and $(n-1)k = (m-1)l$, $\sigma_{st}^2 \leqslant \sigma_r^2$, if the following conditions be satisfied: (i) $\Delta_1\psi \leqslant 0$, (ii) $\Delta_2\psi \leqslant 0$ and (iii) $\Delta_1\Delta_2\psi \geqslant 0$ and $\sigma_{st}^2 < \sigma_r^2$ unless equality holds in each of these three cases.

The parametric relation is, of course, satisfied, if $k = l$ and $m = n$. This theorem will be proved with the help of lemma IV.

Lemma IV. The expression $\sum\limits_{i=0}^{k} \sum\limits_{j=0}^{l} \alpha_{ij}\psi(i,j) \geqslant 0$, if the following conditions be satisfied:

(i) $\sum\limits_{i=0}^{u} \sum\limits_{j=0}^{v} \alpha_{ij} \geqslant 0$ for $u < k$, $v < l$; but $\sum\limits_{i=0}^{k} \sum\limits_{j=0}^{l} \alpha_{ij} = 0$,

(ii) $\Delta_2\psi(k,v) \leqslant 0$, (iii) $\Delta_1\psi(u, l) \leqslant 0$, and (iv) $\Delta_1\Delta_2\psi(u, v) \geqslant 0$.

Proof. $\sum\limits_{i=0}^{k} \sum\limits_{j=0}^{l} \alpha_{ij}\psi(i,j) = -\sum\limits_{j=0}^{l} \sum\limits_{u=0}^{k-1} S_{uj}\Delta_1\psi(u,j) + \sum\limits_{j=0}^{l} S_{kj}\psi(k,j)$, where $S_{uj} = \sum\limits_{i=0}^{u} \alpha_{ij}$

$$= \sum_{u=0}^{k-1} \sum_{v=0}^{l-1} T_{uv}\Delta_1\Delta_2\psi(u,v) - \sum_{u=0}^{k-1} T_{ul}\Delta_1\psi(u,l) - \sum_{v=0}^{l-1} T_{kv}\Delta_2\psi(k,v) + T_{kl}\psi(k,l)$$

where $T_{uv} = \sum\limits_{j=0}^{v} S_{uj} = \sum\limits_{i=0}^{u} \sum\limits_{j=0}^{v} \alpha_{ij}$ whence the lemma can be proved.

Now, we shall prove theorem 3.

$$L(l+1,\ k+1) -- L(l,k) = \sum_{v=1}^{l} \alpha_{0v}\psi(0,v) + \sum_{u=1}^{k} \alpha_{u0}\psi(u,0) + \sum_{u=1}^{k} \sum_{v=1}^{l} \alpha_{uv}\psi(u,v) = \sum_{i=0}^{k} \sum_{j=0}^{l} \alpha_{ij}\psi(i,j)\ ;$$

where putting $k' = k+1$, $l' = l+1$, $A = lk-1$, and $A' = l'k'-1$

$$\alpha_{00} = 0,\ \alpha_{0j} = \frac{l-j}{lA} - \frac{l'-j}{l'A'},\ \alpha_{i0} = \frac{k-i}{kA} - \frac{k'-i}{k'A'}$$

and $$\alpha_{ij} = \frac{2(k-i)(l-j)}{lkA} -- \frac{2(k'-i)(l'-j)}{l'k'A'},\ \text{for } i \neq 0,\ j \neq 0.$$

Then $T_{uv} = \sum\limits_{i=1}^{u} \alpha_{i0} + \sum\limits_{j=1}^{v} \alpha_{0j} + \sum\limits_{i=1}^{u} \sum\limits_{j=1}^{v} \alpha_{ij}$, whence

$$2T_{uv} = \frac{u(k+x-1)}{kA} + \frac{v(l+y-1)}{lA} + \frac{uv(k+x-1)(l+y-1)}{lkA} - \frac{u(k'+x)}{k'A'} - \frac{v(l'+y)}{l'A'} - \frac{uv(k'+x)(l'+y)}{l'k'A}$$

where $x = (k-u)$ and $y = (l-v)$, from which it can be proved that $T_{uv} \geqslant 0$, if $\sqrt{l} \leqslant k \leqslant l^2$, and T_{kl} is always zero ; *i.e.*, if $\sqrt{l} \leqslant k \leqslant l^2$, $\sum\limits_{i=0}^{k} \sum\limits_{j=0}^{l} \alpha_{ij} = 0$

and $\sum\limits_{i=0}^{u} \sum\limits_{j=0}^{v} \alpha_{ij} \geqslant 0$.

Hence, applying lemma IV

$L(l+1,\ k+1) - L(l,k) \geqslant 0$, if (i) $\Delta_1\psi(u,l) \leqslant 0$, (ii) $\Delta_2\psi(u,l) \leqslant 0$ and (iii) $\Delta_1\Delta_2\psi(u,v) \geqslant 0$

whence $L(l, k) \leqslant L(l+1, k+1) \leqslant \ldots, \leqslant L(ml, nk)$ if (i) $\Delta_1\psi \leqslant 0$, (ii) $\Delta_2\psi \leqslant 0$, (iii) $\Delta_1\Delta_2\psi(u,v) \geqslant 0$, (iv) $(m-1)l = (n-1)k$, and (v) $\sqrt{l} \leqslant k \leqslant l^2$.

Hence the theorem.

SECTION 3

We shall now find out a set of sufficient conditions under which the systematic sampling is more efficient than the stratified sampling.

Theorem 4. For all infinite populations, in which (i) $\delta_1^2\, \rho(u,v) \geqslant 0$, and (ii) $\delta_2^2\rho\ (u,v) \geqslant 0$, $\sigma_{sy}^2 \leqslant \sigma_{st}^2$, for any size of the sample and $\sigma_{sy}^2 < \sigma_{st}^2$ unless equality holds in each of the above two cases.

Proof. As the relative efficiency $\sigma_{sy}^2/\sigma_{st}^2$ is independent of ν, we can take it to be unity, *i.e.*, take one sample from each of mn strata.

From (1.13) for $\nu=1$, average $(\bar{x}_{sy} - \bar{x})^2$, the estimate of σ_{sy}^2 is

$$\frac{1}{mnlk}\left\{\sum_{i=1}^{mn}\sum_{j=1}^{lk}(x_{ij}-\bar{x})^2 - \frac{1}{mn}\times(\text{average s.s. within systematic samples})\right\} \quad \ldots \quad (3.1)$$

Let $x_{1j}, x_{2j}, \ldots, x_{mn,\,j}$ be the stratified samples from the mn strata, $\bar{x}_{st}$ the mean ; x_{ij} being the sample from the ith stratum.

Then ${}_j\bar{x}_{st} = \sum_{i=1}^{mn} x_{ij}/mn$. Let $\bar{x}_i$ be the mean of the ith stratum and $\bar{x}$ the grand mean.

Then $\bar{x}_i = \frac{1}{lk}\sum_{j=1}^{lk} x_{ij}$, $\bar{x} = \frac{1}{mn}\sum_{i=1}^{mn}\bar{x}_i$, $\sigma_{st}^2 = \mathrm{E}(\bar{x}_{st}-\bar{x})^2$. Now, $\sum_{i=1}^{mn}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{mn}(x_{ij}-\bar{x}_{st})^2$
$+mn(\bar{x}_{st}-\bar{x})^2$.

Averaging over $(lk)^{mn}$ samples we have

$$\frac{1}{lk}\sum_{i=1}^{mn}\sum_{j=1}^{lk}(x_{ij}-\bar{x})^2 = (\text{average s.s. within sample}) + mn(\bar{x}_{st}-\bar{x})^2.$$

$$\therefore\ \sigma_{st}^2 = \mathrm{E}(\bar{x}_{st}-\bar{x})^2 = \frac{1}{mnlk}\,\mathrm{E}\left\{\sum_{i=1}^{mn}\sum_{j=1}^{lk}(x_{ij}-\bar{x})^2 - \frac{1}{mn}\,(\text{average s.s. within sample})\right\} (3.2)$$

If $x_1, x_2, \ldots, x_{mn}$ be the sample, and $\hat{x}$ its mean, then s.s. within sample is

$$\sum_{i=1}^{mn}(x_i-\hat{x})^2 = \frac{1}{mn}\sum_{\substack{i,j=1\\ j>i}}^{mn}(x_i-x_j)^2\,; \quad \ldots \quad (3.3)$$

which is the average s.s within the sample, systematic or stratified.

Let Σ_{sy} and Σ_{st} be the expectations of the average s.s. within systematic and stratified samples respectively. Then $\sigma_{st}^2 \geqslant \sigma_{sy}^2$, if $\Sigma_{sy} \geqslant \Sigma_{st}$, from (3.1) and (3.2).

From(3.3), Σ_{sy} and Σ_{st} are given by $\frac{1}{mn}$ multiplied by the sum of terms like $E(x_i - x_j)^2$, for all combinations of pairs of strata, (i, j donoting strata).

We can, therefore, compare Σ_{sy} and Σ_{st} term by term.

Let E_{sy} and E_{st} denote expectations for systematic and stratified samples respectively. Also let x_i, x_j belong to (i_1, i_2) and (j_1, j_2) strata respectively, such that $(i_1 - j_1) = u$ and $(i_2 - j_2) = v$.

Then
$$E_{sy}(x_i - x_j)^2 = E_{sy}(x_i - \mu + \mu - x_j)^2 = 2\sigma^2(1 - \rho(uk, vl)), \quad \ldots \quad (3.4)$$

x_j being determined when x_i is fixed and *vice versa.*

For $E_{st}(x_i - x_j)^2$, l^2k^2 combinations of pairs are possible in the present case. We are to take mean over all these combinations.

It can be seen that $(k - |i|)$ pairs are $(ku + i)$ distance apart and $(l - |j|)$ are $(lv + j)$ distance apart.

Averaging over l^2k^2 combinations we have

$$E_{st}(x_i - x_j)^2 = 2\sigma^2 \left\{ 1 - \frac{1}{k^2l^2} \sum_{i=-(k-1)}^{(k-1)} \sum_{j=-(l-1)}^{(l-1)} (k - |i|)(l - |j|)\rho(uk+i, vl+j) \right\} \quad \ldots \quad (3.5)$$

From (3.4) and (3.5), $(\Sigma_{sy} - \Sigma_{st})$ will be proportional to the sum of terms like

$$\frac{1}{k^2l^2} \sum_{i=-(k-1)}^{(k-1)} \sum_{j=-(l-1)}^{(l-1)} (k - |i|)(l - |j|)\rho(uk+i, vl+j) - \rho(uk, vl) = \frac{1}{k^2l^2}(T_1 + T_2 + T_3 + T_4),$$

where,
$$T_1 = \sum_{i=1}^{(k-1)} \sum_{j=1}^{(l-1)} (k - i)(l - j)\{\rho(uk + i, vl + j) + \rho(uk + i, vl - j)$$
$$+ \rho(uk - i, vl + j) + \rho(uk - i, vl - j)\},$$

$$T_2 = k \sum_{j=1}^{(l-1)} (l - j)\{\rho(uk, vl + j) + \rho(uk, vl - j)\}, \; T_3 = l \sum_{i=1}^{k-1} (k - i)\{\rho(uk + i, vl)$$
$$+ \rho(uk - i, vl)\}, \text{ and } T_4 = -lk(lk - 1)\rho(uk, vl).$$

Let us subtract $4\rho(uk, vl)$ from the terms of T_1, and $2\rho(uk, vl)$ from terms of T_2 and T_3 each and add the balance to T_4.

Then we get :
$$T_1' = \sum_{i=1}^{k-1} \sum_{j=1}^{l-1} (k - i)(l - j)\{\rho(uk+i, vl+j) + \rho(uk+i, vl-j) + \rho(uk-i, vl+j)$$
$$+ \rho(uk-i, vl-j) - 4\rho(uk, vl)\}$$

$$T_2' = k\sum_{j=1}^{l-1}(l-j)\{\rho(uk,\ vl+j) + (\rho(uk,\ vl-j)-2\rho(uk,vl)\}, \text{ and}$$

$$T_3' = l\sum_{i=1}^{k-1}(k-i)\{\rho(uk+i,\ vl)+\rho(uk-i,\ vl) - 2\rho(uk,vl)\}$$

We are in all subtracting $lk(lk-1)\,\rho(uk,\ vl)$, and adding the same to T_4, we get zero as resultant.

Hence the expression is $\qquad T_1'+T_2'+T_3' \qquad$ ... (3.6)

If as usual $\delta^2 = \Delta^2E^{-1} = (E^{\frac{1}{2}}-E^{-\frac{1}{2}})^2$, it can be proved that

$$\sum_{j=-(i-1)}^{(i-1)}(i-|j|)\delta^2E^j = E^i+E^{-i}-2\ ;$$

Then
$$\sum_{j=-(i-1)}^{(i-1)}(i-|j|)\delta^2\phi(u+j)=\phi(u+i)+\phi(u-i)-2\phi(u) \qquad \text{... (3.7)}$$

Hence in T_1', $\quad \rho(uk+i,\ vl+j)+\ldots+\rho(uk-i,\ vl-j)-4\rho(uk,vl)$

$$= (E_1^i+E_1^{-i}-2)\{\rho(uk,vl+j)+\rho(uk,vl-j\} + 2(E_2^j+E_2^{-j}-2)\rho(uk,\ vl)$$

$$= \sum_{s=-(i-1)}^{(i-1)}(i-|s|)\{\delta_1^2\rho(uk+s,\ vl+j)+\delta_1^2\,\rho(uk+s,vl-j)\}$$

$$+\ 2\sum_{t=-(j-1)}^{(j-1)}(j-|t|)\delta_2^2\,\rho(uk,vl+t), \text{ using (3.7)}$$

Similarly, in T_2'

$$\rho(uk,\ vl+j)+\rho(uk,\ vl-j)-2\rho(uk,\ vl) = \sum_{t=-(j-1)}^{(j-1)}(j-|t|)\ \delta_2^2\ \rho(uk,\ vl+t),$$

and in T_3'

$$\rho(uk+i,\ vl)+\rho(uk-i,\ vl)-2\rho(uk,vl) = \sum_{s=-(i-1)}^{(i-1)}(i-|s|)\ \delta_1^2\rho(uk+s,\ vl).$$

Thus T_1', T_2' and T_3' are each positive, if (*i*) $\delta_1^2\,\rho(u,v) \geqslant 0$ and (*ii*) $\delta_2^2\rho(u,v) \geqslant 0$ for all values of u and v.

Hence under these conditions $\Sigma_{sy} > \Sigma_{st}$ whence $\sigma_{sy}^2 < \sigma_{st}^2$.

It is evident that the difference vanishes only when equality holds in each of the above three cases. Thus the theorem is proved.

From the proof of the theorems given in this section and in previous section, we can find conditions under which $\sigma_r^2 > \sigma_{sy}^2$. In view of the fact that in certain cases, we are compelled to adopt systematic of stratified types of sampling {technique, we ought to find out the degree of inefficiency also, which, though not easily expressible, can be obtained from (1.14), (1.15) and (1.16) of section I.

Now, we shall consider one example.

Let $\rho(u,v) = \sum_{i=1}^{p} A_i e^{-\lambda_i|u|-\mu_i|v|}$, where $\sum_{i=1}^{p} A_i = 1$, and λ_i and μ_i are positive.

Evidently, $\rho(u,v) = \rho(-u,v)$ and we can consider only positive values of u and v. This is supposed to be a natural condition (Ghosh, 1949).

Here, $\psi(u, v) = 2\rho(u, v)$. Then $\Delta_1 \psi(u, v) = 2 \sum_{i=1}^{p} A_i e^{-\lambda_i u-\mu_i v} (e^{-\lambda_i} -1)$.

But when $\lambda_i < 0, e^{-\lambda_i} < 1$; therefore, $\Delta_1\psi(u, v) < 0$, for positive correlations.

Similar is the case with $\Delta_2\psi(u,v)$.

But without knowing the values of λ_i and μ_i we cannot say whether conditions of theorem I are satisfied or not. Only if $\lambda_i = \mu_i$ for all values of i, the corollary to theorem I is applicable.*

Again $\Delta_1\Delta_2 \psi(u, v) = 2A_i e^{-\lambda_i u-\mu_i v}(e^{-\lambda_i}-1)(e^{-\mu_i}-1)$, which is positive.

Hence, theorem 3 can be applied, i.e., stratified sampling with $\sqrt{l} < k < l^2$ and $(n-1)k = (m+1)l$ is more efficient than random sampling, though the universe is homogeneous as regards the expected means.

Also, $\delta_1^2 \psi(u, v) = 2\{\rho(u+1, v)+\rho(u-1, v)-2\rho(u, v)\}$

$$= 2 \sum_{i=1}^{p} A_i e^{-\lambda_i u-\mu_i v}(e^{\lambda_i/2}-e^{-\lambda_i/2})^2, \text{ which is positive for positive } \rho.$$

*It has been also proved by the author that for $\Delta_1\psi(u, v) < 0$, $\Delta_2\psi(u, v) < 0$, $\delta_1^2\rho(u, v) > 0$ and $\delta_2^2\rho(u, v) > 0$ the relation $\sigma_r^2 > \sigma_{st}^2 > \sigma_{sy}^2$ holds good and $\sigma_r^2 = \sigma_{st}^2$ if equality holds in each of the above four cases when $\sigma_{st}^2 = \sigma_{sy}^2$ also will be true. These results will be published in a forthcoming issue of *Science and Culture*.

Thus in this example $\sigma_r^2 > \sigma_{st}^2 > \sigma_{sy}^2$.

Similarly, $\delta_2^2\psi(u, v) > 0$,

Therefore applying theorem 4, systematic sampling is more efficient than stratified sampling.

While the paper was in press, it came to the notice of the author that some of the results discussed in this paper had been also obtained by Quenonille (1949). The present author published these results in Science and Culture in 1949.

Thanks are due to Prof. B. N. Ghosh for some valuable suggestions.

References

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

2. Das, A. C. (1949) : Two dimensional systematic, stratified and random sampling. *Proc. Ind. Sci. Congress*, 36*th Session*, Part III, 6.

3. Ghosh, B. N. (1949) : On a particular type of natural field. *Proc. Ind. Sci. Congress*, 36*th Session*, Part III, 7.

4. Madow, W. G. and L. H. (1944) : On the theory of systematic sampling, I. *Annals of Mathematical Statistics*, 15, 1-24.

5. Quenouille, M. H. (1949) : Problems in plane sampling. *Annals of Mathematical Statistics*, 20, 355-375.

6. Yates, F (1948) : Systematic sampling. *Phil. Trans, Roy.-Soc.*, 241(A), 345-377.

Paper received : 2 February, 1949.