EFFICIENCY OF SUB-SAMPLING DESIGNS IN YIELD SURVEYS

By P. V. Sukhatme

Indian Council of Agricultural Research, New Delhi

In province-wide yield surveys, covering tens of thousands of square miles, the smallest administrative divisions, which can normally be adopted as the strata, are of the size of about 100 to 200 square miles, such as a taluka or revenue inspector's circle. When, however, the domain of the yield survey is itself of the size of a few hundred square miles or smaller, the question arises whether stratification by patwari circles, which are the smallest administrative divisions of the country consisting of about seven square miles on the average, as in the Delhi Province, with the usual sub-sampling by villages and fields within strata, is efficient and necessary. Equally, the possibility of using a patwari circle itself as a primary sampling unit in place of a village also needs to be explored. It is the object of this paper to set out the relevant sampling theory and illustrate its application on the yield data relating to sample surveys carried out in the Delhi Province during the three years 1946–47, 1947–48 and 1948–49 (Sukhatme and Aggarwal, 1949).

2. The Province of Delhi has an area of 579 square miles. It is divided into 87 circles, each in charge of a patwari, containing roughly between 2 to 11 villages per circle. The sampling plan adopted in the yield surveys was stratified multi-stage sampling, with patwari circles as the strata, a village within a patwari circle as the unit of sampling, a field within a village as the sub-unit of sampling and a plot of 1/80 of an acre as the ultimate unit of sampling. From each stratum, two villages were selected, in each selected village two wheat-growing fields were sampled and in each field so selected, a plot of 1/80 of an acre in size ($33' \times 16\frac{1}{2}'$) was randomly located and harvested. In actual practice, the plan could not be strictly adhered to and there were small variations in the numbers of villages and fields sampled. As only one plot was sampled in each selected field, the plan of sampling for all purposes was one of two-stage sampling. The analysis of variance of plot yields for each of the three years is given below:—

shown that the maximum number of types of these components is ${}^{s-1}H_k$, this upper bound being attained when the $s-1$ unitary components of each main effect are all of different types. When, however, these reduce to only q distinct types, the lower bound for the number of distinct types of components is ${}^{q}H_k$. It is surmised that this lower bound is always equalled when s is a prime number and also sometimes when s is a power of 2.

It is not known whether a method exists by which expressions for the components of a $(k-1)$-th order interaction belonging to one type can be derived from the given expression for one such component.

Finally, it is a pleasure to thank Messrs. R. M. Chatterjee and R. C. Pandya for valuable assistance in the extensive numerical calculations.

REFERENCES

1. Bose, R. C. and Kishen, K. .. "On the problem of confounding in the general symmetrical factorial design," *Sankhya*, 1940, 5, 21–36.

2. Chrystal, G. .. *Algebra : An Elementary Text-book*, Part II, Second Edition. A & C Black, Ltd., London, 1931.

3. Kishen, K. .. "On expressing any single degree of freedom for treatments in an s^m factorial arrangement in terms of its sets for main effects and interactions," *Sankhya*, 1942, 6, 133–40.

TABLE I

Analysis of Variance per Plot of Yield of Wheat

(Ch. per 1/80 of an acre)

Source of variation	1946–47		1947–48		1948–49	
	d.f.	*m sq.*	*d.f.*	*m.sq.*	*d.f.*	*m.sq.*
Beween circles ..	77	1855·9	82	1975·5	81	2411·6 S^2_{pc}
Between villages within circles ..	64	874·4	77	1129·3	78	873·9 S^2_{cv}
Between fields within villages ..	142	435·9	157	616·9	160	756·9 S^2_{vf}
Mean yield ..	..	63·4	..	73·2	..	72·3
Sampling error ..	..	1·85	..	1·97	..	1·81

The last column shows the notation for the mean squares corresponding to the three sources of variation, the suffixes p, c, v and f standing for the province or population, circle, village and field respectively.

3. Let

N be the number of units of sampling in the population;

N_j, the number of units in the j-th stratum;

n, the number of units in the sample;

n_j, the number of units in the sample from the j-th stratum;

M, the number of sub-units in each unit of sampling;

m_{ji}, the number of sub-units sampled from the i-th unit in the j-th stratum; and

k, the number of strata in the population.

Also let

x_{jil} be the value of the character for the l-th sub-unit in the i-th unit of the j-th stratum;

$\bar{x}_{ji(M)}$, the true mean of the i-th unit of the j-th stratum so that

$$\bar{x}_{ji(M)} = \frac{1}{M} \sum_{l=1}^{M} x_{jil} \qquad (1)$$

$\bar{x}_{ji\,(m_{ji})}$, the corresponding sample mean, so that

$$\bar{x}_{ji\,(m_{ji})} = \frac{1}{m_{ji}} \sum_{l=1}^{m_{ji}} x_{jil} \tag{2}$$

$\bar{x}_{j\,(N_j,\,M)}$, the true mean of the j-th stratum, so that

$$\bar{x}_{j\,(N_j,\,M)} = \frac{1}{N_j} \sum_{i=1}^{N_j} \bar{x}_{ji\,(M)} \tag{3}$$

$\bar{x}_{j\,(n_j,\,m_{ji})}$ or simply $\bar{x}_{nj}$, the corresponding unweighted sample mean, so that

$$\bar{x}_{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \bar{x}_{ji\,(m_{ji})} \tag{4}$$

$\bar{x}_{(k,\,N_j,\,M)}$ or $\bar{x}_{NM}$ the population mean, so that

$$\bar{x}_{NM} = \frac{1}{N} \sum_{j=1}^{k} N_j\, \bar{x}_{N_jM} \tag{5}$$

and finally

$\bar{x}_{(k,\,n_j,\,m_{ji})}$ or simply $\bar{x}_n$, the sample mean, so that

$$\bar{x}_n = \frac{1}{N} \sum_{j=1}^{k} N_j\, \bar{x}_{n_j} \tag{6}$$

Then it can be shown that the variance of the sample mean $\bar{x}_{nm}$ when the number of sub-units sampled from each of n selected units is m is given by

$$V(\bar{x}_{nm}) = \sigma^2_{cv} \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) + \sigma^2_{vf} \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j m} - \frac{1}{N_j M}\right) \tag{7}$$

where σ^2_{vf} and $(M\sigma^2_{cv} + \sigma^2_{vf})$ are the mean squares between sub-units within units and between units within strata in the population respectively, being given by

$$\sigma^2_{vf} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{N_j} \sum_{l=1}^{M} (x_{jil} - \bar{x}_{ji\,(M)})^2}{\sum_{j=1}^{k} N_j\,(M-1)} \tag{8}$$

$$M\sigma^2_{cv}+\sigma^2_{vf}=\frac{M\sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{x}_{ji\,(M)}-\bar{x}_{j\,(N_j,\,M)})^2}{\sum_{j=1}^{k}(N_j-1)} \tag{9}$$

When M is large as compared to m so that m/M is negligible (7) can be written as

$$V(\bar{x}_{nm})=\sigma^2_{cv}\sum_{j=1}^{k}\frac{N^2_j}{N^2}\left(\frac{1}{n_j}-\frac{1}{N_j}\right)+\sigma^2_{vf}\sum_{j=1}^{k}\frac{N^2_j}{N^2}\cdot\frac{1}{n_jm} \tag{10}$$

Equation (10) may be used to provide an approximation to the variance of the mean yield in crop surveys where the number of fields in a village, though variable, is large as compared to the number sampled We have the mean yield and its variance given by

$$\bar{x}_{nm}=\frac{1}{A}\sum_{j=1}^{k}A_j\,\bar{x}_{n_jm} \tag{11}$$

and

$$V(\bar{x}_{nm})=\sigma^2_{cv}\sum_{j=1}^{k}\frac{A^2_j}{A^2}\left(\frac{1}{n_j}-\frac{1}{N_j}\right)+\sigma^2_{vf}\sum_{j=1}^{k}\frac{A^2_j}{A^2}\frac{1}{n_jm} \tag{12}$$

where A_j is the area under crop in the j-th stratum, and A is the total area under the crop in the province, the suffixes c, v and f standing for circle, village and field respectively. It can be shown that the unbiassed estimates of σ^2_{cv} and σ^2_{vf} are provided by

$$\hat{\sigma}^2_{cv}=\frac{S^2_{cv}-S^2_{vf}}{m} \tag{13}$$

and

$$\hat{\sigma}^2_{vf}=S^2_{vf} \tag{14}$$

respectively, where as already shown in the analysis of variance Table I, S^2_{cv} and S^2_{vf} denote the mean squares between units within strata and between sub-units within units respectively. When the number of fields sampled from a village is not constant, the estimate of σ^2_{vf} would still be given by the mean square between sub-units within units but that of σ^2_{cv} would be given by

$$\hat{\sigma}^2_{cv}=\frac{S^2_{cv}-S^2_{vf}}{\bar{m}} \tag{15}$$

where

$$\bar{m} = \frac{1}{n-k} \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} m_{ji} - \frac{\sum_{i=1}^{n_j} m_{ji}^2}{\sum_{i=1}^{n_j} m_{ji}} \right) \tag{16}$$

m_{ji} $(i = 1, 2, \ldots n_j)$ denoting the number of sub-units sampled in the i-th sampled unit of the j-th stratum (Cochran, 1939).

5. We shall discuss the efficiency of stratification in terms of the relative magnitudes of the variances of the estimated mean (*a*) with strata and (*b*) without strata. An examination of the formula in (7) for the variance of the mean in a sub-sampling design will show that all that is needed in estimating the variance of the mean, appropriate for sampling of units and sub-units without stratification, is an estimate of the quantity σ^2_{pv} representing the variation between true unit means in the population (without strata) in contrast to σ^2_{cv} standing for variation between units within circles (strata) in the population. It follows that the gain due to stratification is simply the percentage increase of the variance given by

$$V(\bar{x}_{nm}) = \sigma^2_{pv}\left(\frac{1}{n} - \frac{1}{N}\right) + \sigma^2_{vf}\left(\frac{1}{nm} - \frac{1}{NM}\right) \tag{17}$$

over the variance given by (7). The effect of varying the numbers of units and sub-units per unit on the gains due to stratification is also easily calculated. For any hypothetical values of n_j's and m, say n_j' and m', the variance of the mean with strata and without strata will be

$$V_{st}(\bar{x}_{n'm'}) = \sigma^2_{cv} \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j'} - \frac{1}{N_j}\right) + \sigma^2_{vf} \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j' \cdot m'} - \frac{1}{N_j M}\right) \tag{18}$$

and

$$V_R(\bar{x}_{n'm'}) = \sigma^2_{pv}\left(\frac{1}{n'} - \frac{1}{N}\right) + \sigma^2_{vf}\left(\frac{1}{n'm'} - \frac{1}{NM}\right) \tag{19}$$

where σ^2_{pv}, σ^2_{cv} and σ^2_{vf} are estimated from the observed sample of n_j units and m_{ji} $(i = 1, 2, \ldots n_j)$ sub-units in the i-th sampled unit in the j-th stratum. The gain in stratification for given n' and m' is the percentage increase of (19) over (18). Since σ^2_{vf} and σ^2_{cv} can be estimated from (14) and (15), the problem of evaluating the

efficiency of stratification in a sub-sampling design thus reduces to the estimation of σ^2_{pv} where on the analogy of (9)

$$\sigma^2_{pv} = \frac{\sum_{i=1}^{N} (\bar{x}_{i\,(M)} - \bar{x}_{NM})^2}{N-1} - \frac{\sigma^2_{vf}}{M} \tag{20}$$

6. In order to simplify the algebra, let us denote the true unit mean $\bar{x}_{i\,(M)}$ by y_i so that

$$\left.\begin{aligned} \bar{x}_{ji\,(M)} &= y_{ji} \\ \bar{x}_{j\,(N_jM)} &= \bar{y}_{N_j} = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ji} \\ \bar{x}_{NM} &= \bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{j=1}^{k} N_j \bar{y}_{N_j} \end{aligned}\right\} \tag{21}$$

and

$$\sigma^2_{pv} = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_N)^2}{N-1} - \frac{\sigma^2_{vf}}{M}$$

Now

$$y_{ji} - \bar{y}_N = y_{ji} - \bar{y}_{Nj} + \bar{y}_{Nj} - \bar{y}_N \tag{22}$$

so that squaring and summing over all the N units

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j} (y_{ji} - \bar{y}_N)^2 = \sum_{i=1}^{N} (y_i - \bar{y}_N)^2 = \sum_{j=1}^{k}\sum_{i=1}^{N_j} (y_{ji} - \bar{y}_{N_j})^2 + \sum_{j=1}^{k} N_j \bar{y}^2_{N_j} - N\bar{y}^2_N \tag{23}$$

Denoting $y_{ji} - \bar{y}_{N_j}$ by ϵ_{ji} so that

$$\sum_{i=1}^{N_j} \epsilon_{ji} = N_j \bar{\epsilon}_{N_j} = 0 \text{ and } \frac{1}{n_j}\sum_{i=1}^{n_j} \epsilon_{ji} = \bar{\epsilon}_{n_j}$$

ve have

$$y_{ji} = \epsilon_{ji} + \bar{y}_{N_j} \tag{24}$$

$$\bar{y}_{n_j} = \bar{\epsilon}_{n_j} + \bar{y}_{N_j} \tag{25}$$

Squaring and taking the expected values of both sides in (25) and remembering that

$$E(\bar{\epsilon}^2_{n_j}) = \frac{N_j - n_j}{N_j}\left(\sigma^2_{cv} + \frac{\sigma^2_{vf}}{M}\right)\frac{1}{n_j} \tag{26}$$

we have on multiplying by N_j and summing over all strata

$$\sum_{j=1}^{k} N_j \bar{y}^2_{N_j} = E\left(\sum_{j=1}^{k} N_j \bar{y}^2_{n_j}\right) - \left(\sigma^2_{cv} + \frac{\sigma^2_{vf}}{M}\right)\left(\sum_{j=1}^{k} \frac{N_j - n_j}{n_j}\right) \tag{27}$$

From (25)

$$\sum_{j=1}^{k} N_j \bar{y}_{n_j} = \sum_{j=1}^{k} N_j \bar{\epsilon}_{n_j} + \sum_{j=1}^{k} N_j \bar{y}_{N_j}$$

Whence

$$N\bar{y}_n = \sum_{j=1}^{k} N_j \bar{\epsilon}_{n_j} + N\bar{y}_N \tag{28}$$

Squaring and taking the expected values, dividing by N and transferring terms

$$N\bar{y}^2_N = E(N\bar{y}^2_n) - \frac{1}{N}\left(\sigma^2_{cv} + \frac{\sigma^2_{vf}}{M}\right)\sum_{j=1}^{k} N_j \left(\frac{N_j - n_j}{n_j}\right) \tag{29}$$

Substituting the results from (9), (27) and (29) in (23), we obtain

$$\sum_{i=1}^{N} (y_i - \bar{y}_N)^2 = \left(\sigma^2_{cv} + \frac{\sigma^2_{vf}}{M}\right)\left\{(N-k) - \sum_{j=1}^{k}\left(\frac{N_j - n_j}{n_j}\right)\left(1 - \frac{N_j}{N}\right)\right\} + E\left\{\sum_{j=1}^{k} N_j \bar{y}^2_{n_j} - N\bar{y}^2_n\right\} \tag{30}$$

We now proceed to obtain for a fixed n_j the estimate of the last two terms in (30), namely,

$$\sum_{j=1}^{k} N\bar{y}^2_{n_j} - N\bar{y}^2_n = \sum_{j=1}^{k} N_j \bar{x}^2_{n_{j(M)}} - N\bar{x}^2_{n(M)} \tag{31}$$

Now

$$x_{jil} = x_{jil} - \bar{x}_{ji(M)} + \bar{x}_{ji(M)}$$

$$= \epsilon_{jil} + \bar{x}_{ji(M)}$$

whence taking the mean over the m_{ji} sub-units in the i-th unit of the j-th stratum

$$\bar{x}_{ji(m_{ji})} = \bar{\epsilon}_{ji(m_{ji})} + \bar{x}_{ji(M)} \tag{33}$$

Summing (33) over the n_j units in j-th stratum and taking the mean we have

$$\bar{x}_{j\,(n_j,\,m_{ji})} = \bar{x}_{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \bar{\epsilon}_{ji\,(m_{ji})} + \bar{x}_{n_j\,(M)} \tag{34}$$

Squaring and taking the expectations for a fixed n_j, we obtain

$$E(\bar{x}_{n_j}^2) = \frac{1}{n_j^2} \left\{ \sum_{i=1}^{n_j} \frac{M - m_{ji}}{M} \frac{\sigma_{vj}^2}{m_{ji}} \right\} + \bar{x}^2_{n_j\,(M)} \tag{35}$$

whence

$$\sum_{j=1}^{k} N_j \bar{x}^2_{n_j\,(M)} = E \left\{ \sum_{j=1}^{k} N_j \bar{x}^2_{n_j} \right\} - \sigma^2_{vf} \sum_{j=1}^{k} \left\{ \frac{N_j}{n_j^2} \sum_{i=1}^{n_j} \frac{M - m_{ji}}{M} \cdot \frac{1}{m_{ji}} \right\} \tag{36}$$

Also from (34) we obtain

$$\sum_{j=1}^{k} N_j \bar{x}_{j\,(n_j,\,m_{ji})} = \sum_{j=1}^{k} N_j \bar{x}_{n_j} = \sum_{j=1}^{k} N_j \bar{\epsilon}_{n_j} + \sum_{j=1}^{k} N_j \bar{x}_{n_j(M)}$$

or

$$N\bar{x}_n = \sum_{j=1}^{k} N_j\, \bar{\epsilon}_{n_j} + N\bar{x}_{n_{(M)}} \tag{37}$$

Squaring and taking expectations, and transferring terms we get

$$N\bar{x}^2_{n\,(M)} = E(N\bar{x}^2_n) - \frac{1}{N} \sum_{j=1}^{k} \frac{N^2_j}{n^2_j} \sum_{i=1}^{n_j} \frac{M - m_{ji}}{M} \frac{\sigma^2_{vf}}{m_{ji}} \tag{38}$$

Hence, substituting in (31) from (36) and (38), we have for the estimate of the expression in (31) the following

$$\sum_{j=1}^{k} N_j\, \bar{x}^2_{n_j} - N\bar{x}^2_n - S^2_{vf} \sum_{j=1}^{k} \frac{N_j}{n^2_j} \left(1 - \frac{N_j}{N}\right) \left(\sum_{i=1}^{n_j} \frac{M - m_{ji}}{M} \cdot \frac{1}{m_{ji}} \right) \tag{39}$$

Substituting in (30), we obtain

$$\hat{\sigma}^2_{pv} = \frac{S^2_{cv} - S^2_{vf}}{\bar{m}} - \frac{1}{N-1}\cdot\frac{S^2_{cv} - S^2_{vf}}{\bar{m}} \sum_{j=1}^{k} \frac{N_j}{n_j}\left(1 - \frac{N_j}{N}\right)$$

$$- \frac{S^2_{vf}}{N-1} \sum_{j=1}^{k} \frac{N_j}{n_j}\left(1 - \frac{N_j}{N}\right)\frac{1}{h_j} + \frac{1}{N-1}\left\{\sum_{j=1}^{k} N_j \bar{x}^2_{n_j} - N\bar{x}^2_n\right\} \tag{40}$$

where

$$\frac{1}{h_j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\frac{1}{m_{ji}}$$

For sake of brevity, we may write (40) as

$$\hat{\sigma}^2_{pv} = \alpha S^2_{cv} + \beta S^2_{vf} + \frac{1}{N-1}\left(\sum_{j=1}^{k} N_j \bar{x}^2_{n_j} - N\bar{x}^2_n\right) \tag{41}$$

where

$$\alpha = \frac{1}{\bar{m}}\left\{1 - \frac{1}{N-1}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1 - \frac{N_j}{N}\right)\right\} \tag{42}$$

$$\beta = \frac{1}{\bar{m}}\left\{-1 + \frac{1}{N-1}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1 - \frac{N_j}{N}\right)\right.$$

$$\left. - \frac{1}{N-1}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1 - \frac{N_j}{N}\right)\frac{\bar{m}}{h_j}\right\} \tag{43}$$

Whence we obtain for the estimate of the variance of the mean without strata

$$V_R(\bar{x}_{n'm'}) = \left\{\alpha S^2_{cv} + \beta S^2_{vf} + \frac{1}{N-1}\left(\sum_{j=1}^{k} N_j \bar{x}^2_{n_j} - N\bar{x}^2_n\right)\right\} \times \left(\frac{1}{n'} - \frac{1}{N}\right)$$

$$+ S^2_{vf}\left(\frac{1}{n'm'} - \frac{1}{NM}\right) \tag{44}$$

7. Table II shows the calculations relating to the gains in stratification on Delhi data for the years 1946–47, 1947–48 and 1948–49. Row No. 1 of the table gives the sampling variance of the observed mean calculated from (12); row No. 2 gives the estimated variance of the mean if villages had been selected at random from the whole population, that is, without strata, calculated from (19); row No. 3

TABLE II

Efficiency of Stratification

Variance in $(chh)^2$	1946–47 m			1947–48 m			1948–49 m		
	1	2	3	1	2	3	1	2	3
1. Stratified sampling (n_j' actual)	5·67	3·41	2·66	6·73	3·89	2·95	6·37	3·28	2·25
2. Unrestricted sampling	5·40	3·87	3·35	5·38	3·45	2·81	6·31	3·94	3·15
3. % Gain in precision	−4·8	13·5	25·9	−20·1	−11·3	−4·7	−0·9	20·1	40·0
4. Stratified sampling ($n_j' \propto A_j$)	3·93	2·25	1·69	4·88	2·69	1·96	5·23	2·66	1·80
5. % Gain in precision	37·4	72·2	98·2	10·2	28·3	43·4	20·6	48·1	75·0

shows the percentage gain or loss in precision; row No. 4 shows the estimated variance if villages had been distributed in proportion to the acreage under wheat in the different patwari circles, taking the numbers of villages to be the nearest whole numbers proportional to acreage and row No. 5 shows the corresponding percentage gain in precision. The effect of varying m' is also brought out in the table. It will be seen that while the actual gains made are small, there is a very considerable scope for increasing the gains in precision by distributing villages in proportion to the acreage under wheat in the different strata.

8. The formula (44) is a general formula for evaluating gains due to stratification in a sub-sampling design. Certain particular cases of interest of this formula are discussed below:

(*a*) *No sub-sampling.*—The appropriate formula is readily derived from (44) by replacing m_{ji} by M.

We have

$$\alpha = \frac{1}{M}\left\{1 - \frac{1}{N-1}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1 - \frac{N_j}{N}\right)\right\}$$

and

$$\beta = -\frac{1}{M}$$

giving

$$V_R(\bar{x}_{n'M}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\left\{\frac{S^2_{cv}}{M} - \frac{S^2_{cv}}{M(N-1)}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1-\frac{N_j}{N}\right) + \frac{1}{N-1}\left(\sum_{j=1}^{k} N_j\bar{x}^2_{n_jM} - N\bar{x}^2_{nM}\right)\right\} \quad (45)$$

Yates (1949) has given an expression appropriate for this case which, however, appears to be approximate and can be derived from (45) by replacing $N-1$ by N wherever it occurs. Yates' expression corresponds to the one given by the author elsewhere (1950).

(*b*) *Sub-sampling when* $m_{ji} = m$ *for all* j *and* i.—The coefficients α and β assume the values

$$\alpha = \frac{1}{m}\left\{1 - \frac{1}{N-1}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1-\frac{N_j}{N}\right)\right\}.$$

$$\beta = -\frac{1}{m}$$

giving

$$V_R(\bar{x}_{n'm'}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\left\{\hat{\sigma}^2_{cv} - \frac{S^2_{cv}}{m(N-1)}\sum_{j=1}^{k}\frac{N_j}{n_j}\left(1-\frac{N_j}{N}\right) + \frac{1}{N-1}\left(\sum_{j=1}^{k} N_j\bar{x}^2_{n_j} - N\bar{x}^2_{n}\right)\right\} + \left(\frac{1}{n'm'} - \frac{1}{NM}\right)S^2_{vf} \quad (46)$$

(*c*) *Sub-sampling when* $m_{ji} = m$ *and* $n_j \propto N_j$.—The expression (44) takes the form

$$V_R(\bar{x}_{n'm'}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\left\{\hat{\sigma}^2_{cv} + \frac{N}{N-1}\frac{k-1}{nm}(S^2_{pc} - S^2_{cv})\right\} + \left(\frac{1}{n'm'} - \frac{1}{NM}\right)S^2_{vf} \quad (47)$$

(*d*) *Sub-sampling when* $m_{ji} = m$, $n_j = \frac{n}{k}$ *and* N_j *is large*.—The appropriate expression is given by

$$V_R(\bar{x}_{n'm'}) = \frac{k-1}{k}\frac{\hat{\sigma}^2_{pc}}{n'} + \frac{\hat{\sigma}^2_{cv}}{n'} + \frac{\hat{\sigma}^2_{vf}}{n'm'} \quad (48)$$

whence the gain in precision when m' fields are sampled per village is simply

$$\frac{\frac{k-1}{k}\hat{\sigma}^2_{pc}}{\hat{\sigma}^2_{cv}+\frac{\hat{\sigma}^2_{vf}}{m'}} \tag{49}$$

This expression is appropriate to the case illustrated by Cochran (1939) with the help of a numerical example for estimating gains due to stratification in a sub-sampling design. Its use in place of (44) gives an exaggerated idea of the efficacy of stratification as will be seen from the following table showing for Delhi data the actual gains calculated from (44) and those calculated from (49):

% gain due to stratification	1946–47	1947–48	1948–49
From (44) ..	13	−11	20
From (49) ..	60	38	55

It is seen that whereas the actual gains made are small ranging from −11 to 20 per cent., the use of formula (49) gives gains which are very much bigger ranging from 38 to 60 per cent.

(*c*) A method sometimes used for estimating gains due to stratification is to obtain a pooled mean square between villages from the mean squares due to strata and villages and compare it with the mean square between villages in the analysis of variance table. This method overestimates the gain even further. For, the variance of the mean of *nm* fields based on the pooled mean square will be

$$\frac{1}{nm}\frac{(k-1)S^2_{pc}+(n-k)S^2_{cv}}{n-1}$$

and exceeds the variance obtained from (48) by

$$\frac{k-1}{nm}\left(\frac{1}{n-1}-\frac{1}{n}\right)(S^2_{pc}-S^2_{cv})$$

9. The results given in the preceding sections also enable us to determine the effect on sampling variance of the alternative methods of using a patwari circle and village as units of sampling. The consideration of this problem is of special importance to the large-scale yield surveys covering tens of thousands of square miles where

sampling every patwari circle, as in Delhi, is out of question and only a fraction of the total number of primary units, amounting at best to only 5 per cent. of the population, can be chosen for the observations. Any conclusion, therefore, which can be reached on Delhi data on the relative efficiency of the two units will be of considerable help in improving the sampling plans for large-scale surveys.

It is easy to see that the variance of the sample mean based on n' circles and m' fields is approximately given by

$$\sigma^2_{pc}\left(\frac{1}{n'}-\frac{1}{k}\right)+\sigma^2_{wf}\left(\frac{1}{n'm'}\right) \tag{50}$$

while that based on n' villages and m' fields is known to be

$$\sigma^2_{pv}\left(\frac{1}{n'}-\frac{1}{N}\right)+\sigma^2_{vf}\left(\frac{1}{n'm'}\right). \tag{51}$$

Consequently the gain in precision by using a circle as unit in place of a village for a given number of experiments is given by the ratio of (50) to (51).

Table III shows the values of the sampling variances calculated from (50) and (51) together with the percentage gain in precision for $n'=5$, 10, 20, 40 and k and $m'=2$. It will be seen from the table that the gain in precision increases with the number of units and ranges from 60 to 120 per cent. when all the patwari circles are sampled. When the number of units is 5, which is about the number usually selected per stratum in large-scale yield surveys, the percentage gain is much smaller though even here it reaches about 25 per cent. Whether this gain will compensate the additional labour involved in sampling fields directly from the circles is a question which deserves to be studied further by undertaking suitable local investigations.

An alternative approach to the study of this problem, which in some ways appears to be more instructive than the one presented above has been suggested by Hansen and Hurwitz (1942). It is based on the concept of intra-class correlation among sub-units of a unit of sampling. If ρ_1 is used to denote the intra-class correlation between sub-units within a village and ρ_2 that between sub-units within a circle, then it can be shown that neglecting terms in $1/M$

$$\rho_1=\frac{\frac{N-1}{N}\sigma^2_{pv}}{\sigma^2_1}$$

6

TABLE III

Relative Efficiency of the Patwari Circle and Village as Units of Sampling with Fields as Sub-units

n	1946–47					1947–48					1948–49				
	5	10	20	40	k	5	10	20	40	k	5	10	20	40	k
Variance with village as the unit	157·1	77·7	38·0	18·1	8·6	151·6	75·1	36·9	17·9	7·9	160·3	84·5	41·5	20·05	9·0
Variance with circle as the unit	124·8	60·2	27·9	11·7	3·9	138·5	67·4	31·9	14·1	4·9	130·0	63·1	30·0	13·4	4·9
% Gain ..	25·9	29·0	36·2	54·3	122·2	9·4	11·4	15·7	26·0	60·8	23·7	33·7	38·5	49·7	85·1
ρ_1 ..		0·56					0·42					0·39			
ρ_2 ..		0·36					0·27					0·25			

where

$$\sigma^2{}_1 \simeq \frac{N-1}{N}\,\sigma^2{}_{pv} + \sigma^2{}_{vf} \tag{52}$$

and

$$\rho_2 = \frac{\frac{k-1}{k}\,\sigma^2{}_{pc}}{\sigma^2{}_2}$$

where

$$\sigma^2{}_2 \simeq \sigma^2 = \frac{k-1}{k}\,\sigma^2{}_{pc} + \sigma^2{}_{cf} \tag{53}$$

The values of ρ_1 and ρ_2 are shown in the last two rows of the table. They show that fields within a village or circle are positively correlated with regard to yield per acre. They further show that correlation decreases with increase in the size of the primary unit (cluster), being approximately two-thirds in a circle as compared to that in a village, indicating greater internal heterogeneity in a circle than in a village. The influence of the correlation on the sampling variance is brought out by substituting the results from (52) and (53), the variances being now given by

$$V(\bar{x}_{n'm'}) = \frac{\sigma^2}{n'm'}\left\{1 + \rho_1\left(\frac{N-n'}{N-1}\,.\,m' - 1\right)\right\} \tag{54}$$

and

$$V(\bar{x}_{n'm'}) = \frac{\sigma^2}{n'm'}\left\{1 + \rho_2\left(\frac{k-n'}{k-1}\,.\,m' - 1\right)\right\} \tag{55}$$

Following Hansen and Hurwitz, the first factor in these formulæ represents the variance of the mean based on $n'm'$ sub-units sampled at random from the whole population, while the second factor measures the contribution to the sampling variance of cluster sampling, the relative change in precision being approximately given by

$$\frac{1 + \rho_1\left(\frac{N-n'}{N-1}\,.\,m' - 1\right)}{1 + \rho_2\left(\frac{k-n'}{k-1}\,.\,m' - 1\right)}. \tag{56}$$

The effect of varying m' on the relative precision has not been shown in the table but can be readily calculated. It is found that the percentage gain increases with m'. Thus for 1946–47 and $n' = 5$, the percentage gain in precision works out to 8 when $m' = 1$, 26 when $m' = 2$ and 37 when $m' = 3$, the gains observed in other years being of

similar order. The corresponding figures for $n' = 40$ are naturally larger, being 22, 54 and 77 respectively.

10. When every Patwari Circle is sampled, the problem of estimating the extent to which the precision of the estimated mean can be improved by sampling fields in one stage, *i.e.*, directly from a circle instead of in two stages, *i.e.*, sampling villages in the first instance and fields in the selected villages, assumes importance. The results of this study when the number of fields sampled is $2k$ are set out in Table IV. Three alternative procedures of distributing

TABLE IV

Relative Efficiency of One versus *Two Stage Sampling in Stratified Sampling with Circles as Strata*

Variance in chh^2	1946–47	1947–48	1948–49
(*a*) One-stage sampling $m \propto A_j$ ..	3·87	4·90	4·89
(*b*) Two-stage sampling—			
(i) $n_j = 1$ and $m = 2$ ..	6·54	8·12	6·73
(ii) $n_j = 2, m = 1$ ..	4·62	5·97	6·25
(iii) $n_j \propto A_j, m = 1$ ..	3·93	4·88	5·23
% Gain in efficiency of (*a*) over *b*(ii) ..	19·3	21·8	27·8
% Gain in efficiency of (*a*) over *b* (iii) ..	1·6	0·0	7·0

experiments as between villages and fields have been considered. Row No. 1 of the table shows the variance of the estimated mean when the number of fields sampled from a circle is proportional to the area under the crop in the circle (method *a*); row No. 2 shows the variance of the estimated mean when one village is sampled from each circle and two fields from each selected village [method *b* (i)]; row No. 3 shows the results when two villages are sampled from each circle and one field is sampled from each village [method *b* (ii)] and row No. 4 shows the variance when the number of villages sampled from each circle is proportional to the area under the crop in the circle and one field sampled from each selected village [method *b* (iii)], the totality of fields remaining the same in all the cases. Rows Nos. 5 and 6 show the estimated gain in precision of method *a* over *b* (ii) and over *b* (iii) respectively. It will be seen that sampling fields directly from a circle

results in a gain of about 20 per cent. over the design in which two villages are sampled from each circle and one field in each selected village. The gains, however, almost disappear when the number of villages sampled from each circle is proportional to the area under the crop in the circle and one field is sampled from each selected village. When every circle has to be sampled, it would, therefore, appear that the plan of sampling, in which villages are distributed in proportion to the area under the different circles and one field is sampled from each selected village, is likely to be both efficient and convenient.

SUMMARY

A general formula appropriate for the estimation of gains in precision due to stratification in a sub-sampling design from finite population has been developed and illustrated on the yield data relating to sample surveys carried out in Delhi Province during 1946–47, 1947–48 and 1948–49. Formulæ appropriate for (*a*) no sub-sampling and (*b*) sub-sampling with a uniform sampling fraction at the first stage are shown to be particular cases of the general formula. Based on the same results, an approach has been indicated for calculating the relative efficiency of sampling units of different size and of one *versus* two stage sampling and the method illustrated on the yield data for Delhi Province.

REFERENCES

Sukhatme, P. V. and Aggarwal, O. P. .. "Report on Random Sample Surveys for Estimating Yield of Wheat in Delhi Province," *Ind. Council of Agri. Research*, 1949.

Sukhatme, P. V. .. "Sample Surveys in Agriculture" *Presidential Address to the Section of Statistics, 37th Indian Science Congress, Poona*, 1950.

Cochran, W. G. .. "The Use of Analysis of Variance in Enumeration by Sampling," *Jour. Amer. Stat. Asso.*, 1939, **34**.

Yates, F. .. *Sampling Methods for Censuses and Surveys*—Charles Griffin & Co. Ltd., 1949.

Hansen, M. H. and Hurwitz, W. N. .. "Relative Efficiencies of Various Sampling Units in Population Inquiries," *Jour. Amer. Stat. Asso.*, 1942, **37**.