SAMPLING WITH REPLACEMENT*

By P. V. Sukhatme and R. D. Narain
Indian Council of Agricultural Research, New Delhi

The Indian Council of Agricultural Research has been conducting Statewide sample surveys for estimating the area and yield of principal crops for the past several years. An account of the yield surveys has already appeared in the previous issues of this *Journal*. The plan of sampling, it will be recalled, consists of stratified multi-stage random sampling with tehsils as the strata, a village as the primary sampling unit, field (a survey number) as the sub-unit of sampling and a plot of a specified size, usually of about 1/80th of an acre as the ultimate unit of sampling. Three fields are selected from each selected village and one plot from each selected field. Equal chance is given to all units of sampling at each stage of selection. The plan of sampling for area estimation has been developed on similar lines. A village constitutes the p.s.u. and a cluster of survey numbers as the sub-unit of sampling, clusters being formed by grouping consecutive survey numbers 1–8, 9–15, etc. (the last cluster consists of the remainder left on division by 8). The sampling procedure for the selection of the p.s.u.'s and the determination of the number of sub-units to be sampled from each selected p.s.u., however, differ from that of the yield survey. The p.s.u.'s are selected with replacement with probabilities proportional to their sizes as measured by the number of sub-units in the village, but the sub-units within the selected p.s.u.'s are selected in the same way as in the yield survey with equal probability without replacement. The number of sub-units to be selected from a selected p.s.u. is determined by the number of times the p.s.u. appears in the sample. Thus, if a village happens to be selected λ times, a sub-sample of $m\lambda$ clusters is selected from it where m is some integer. The field work consists in identifying the selected clusters with the help of the maps and noting the name of the crops grown thereon together with the area under each crop.

This plan of sampling was tried out on a pilot scale in Delhi during 1949–50 and 1950–51 and extended on a Statewide scale in Orissa during the winter of 1950–51. The results of these surveys have shown that with a sampling fraction of less than 1 in 10 thousand the acreage under

* Read at the Fourth Annual Meeting of the Society held in November 1950.

4. International Training Centre on Censuses and Statistics for South-East Asia, "Report on socio-economic survey in Delhi villages" (unpublished), 1950.

5. Indian Council of Agricultural Research, New Delhi, "Report on spot check of patwari's records conducted in Lucknow District," 1950.

6. Indian Council of Agricultural Research, New Delhi, "Tables relating to the agricultural survey conducted by students of the Council in villages of Delhi Development Scheme," (unpublished).

7. Mahalanobis, P. C., "On Large-scale sample surveys," *Phil. Trans. Royal Soc. London*, 1944, **231**, 41.

8. ————————, "Bihar Crop Survey," *Sankhya*, 1945, **7**, 29.

9. U. N. Sub-Commission on Sampling, "Recommendations on the preparation of reports of Sample Surveys" 1948.

10. Sukhatme, P. V. and Panse, V. G., "Crop Surveys in India," *Jour. Ind. Soc. Agr. Stat.*, 1948, **1**, 34.

11. Ghosh, B. N., *Bulletin Calcutta Stat. Asson.*, 1949, **2**, 7.

12. Yates, F., *Sampling Methods for Censuses and Sample Surveys*, Charles, Giffin & Co., London, 1949.

13. Mokashi, V. K., *Journal Ind. Soc. Agr. Stat.*, 1950, **2**, 2.

principal crops can be estimated with standard error approaching 5 per cent. for each district and less than 2 per cent. for the State as a whole. An account of these surveys will be found in the reports of the Indian Council of Agricultural Research.

The sampling theory appropriate to this plan presents a number of interesting features which to our knowledge have not been discussed elsewhere before. It is the object of this paper to present the theory, compare the efficiency of the plan with that of an analogous plan of sampling in which the sub-sampling units are selected independently every time the particular p.s.u. occurs in the sample and also indicate extension of the results to cover a system of sampling in which p.s.u.'s are sampled with replacement until a given number of distinct p.s.u.'s occur in the sample.

2. For sake of simplicity we shall confine the discussion to the case of one single stratum, the results being easily extendable to any number of strata, and adopt the following notation:

N = number of p.s.u.'s (villages) in the stratum.

M_i = number of sub-units (clusters of survey numbers) in the i-th p.s.u.

$M = N\bar{M} = \sum_{i=1}^{N} M_i$.

y_{ij} = the value of j-th sub-unit in the i-th primary unit.

$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$ = population value of the mean per sub-unit in the i-th p.s.u.

$\bar{Y} = \frac{1}{N\bar{M}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$.

$= \frac{1}{N\bar{M}} \sum_{i=1}^{N} M_i \bar{Y}_i$ = mean per sub-unit in the population.

n = number of p.s.u.'s to be selected for the sample with replacement.

v = number of distinct p.s.u.'s in the sample.

λ_i = number of times the i-th selected p.s.u. is repeated in the sample.

$m\lambda_i$ = number of sub-units to be sub-sampled from a selected p.s.u. so that

$$\sum_{i=1}^{v} \lambda_i = n.$$

$$\bar{y}_i = \frac{\sum_{j=1}^{m\lambda_i} y_{ij}}{m\lambda_i} = \text{mean of the sample observations in the } i\text{-th selected unit.}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{\nu} \lambda_i \bar{y}_i = \text{sample mean.}$$

$$\sigma_i^2 = \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2/M_i - 1.$$

$$\sigma_w^2 = \sum_{i=1}^{N} \sigma_i^2 \cdot \frac{M_i}{M} \text{ and } \sigma_b^2 = \sum_{i=1}^{N} \frac{M_i}{M} (\bar{Y}_i - \bar{Y})^2$$

If the probability of drawing the i-th p.s.u. $p_i = \frac{M_i}{M}$, it is easy to see that the sample mean $\bar{y}$ is an unbiassed estimate of the population mean $\bar{Y}$. To obtain the sampling variance, and higher moments if necessary, we make use of the following result:

$$E\left(\sum_{i=1}^{\nu} \lambda_i^r z_i\right) = \sum_{i=1}^{N} z_i \times [r\text{-th moment of the binomial } (n, p_i)].$$

Thus the variance of $\bar{y}$ readily works out to

$$V_1 = \frac{1}{n}\left\{\sigma_b^2 + \sum_{i=1}^{N} \frac{M_i - m}{m\ M_i}\ p_i\sigma_i^2 - (n-1)\sum_{i=1}^{N} \frac{p_i^2\sigma_i^2}{M_i}\right\}.$$

In case sub-sampling within a primary unit is carried out independently every time it is selected, estimate of $\bar{Y}$ would be

$$\bar{y}' = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} \Big/ mn$$

and its variance would be, as is well known,

$$V_2 = \frac{1}{n}\left\{\sigma_b^2 + \sum_{i=1}^{N} \frac{M_i - m}{m\ M_i}\ p_i\sigma_i^2\right\}.$$

In our plan of sampling the part of the variance attributable to sub-sampling is thus reduced to the order of

$$\frac{m\,(n-1)}{(M-m)\,N}\%.$$

which, it may be noted, is very nearly equal to the over-all sampling fraction.

What is perhaps more important is to note that even for large M_i, the estimates of variance obtainable from a sample will differ in

the two cases. In case II when the sub-sampling within a p.s.u. is carried out independently every time it is selected, this is derivable in the usual manner from the analysis of variance table of n units each containing m sub-units. For case I which is the plan considered in this paper, we have only v p.s.u.'s, the i-th unit containing $m\lambda_i$ sub-units, λ_i and v both being random variables. The estimates of σ_b^2 and σ_ω^2 are in this case given by

$$\hat{\sigma}_\omega^2 = \sum_{i=1}^{v} \sum_{j=1}^{m\lambda_i} (\bar{y}_{ij} - y_i)^2/nm - v$$

$$\hat{\sigma}_b^2 = \sum_{i=1}^{v} \frac{\lambda_i (\bar{y}_i - \bar{y})^2}{n-1} - \frac{\hat{\sigma}_\omega^2}{m}\left\{\frac{E(v)-1}{n-1}\right\},$$

where

$E(v)$ = expected number of distinct units

$$= N - \sum_{i=1}^{N} (1-p_i)^n.$$

The data of yield surveys on Wheat in Delhi during 1949–50, 1950–51, conducted according to this plan, were analysed for the purposes of illustration. The original 87 patwari circles in the State were grouped into 59 groups which formed the (k) strata.

Let for the i-th stratum n_i be the number of villages for the sample, v_i the number of distinct villages, and λ_{ij} the number of times j-th villages is repeated. The number of plots in each selected village was two if selected once and $2\lambda_{ij}$ if j-th village of the i-th stratum was selected λ_{ij} times.

The pooled analysis of variance tables relevant to the two plans of sampling are given below:

Analysis of Variance Table (Case I)

(Ch. per $\frac{1}{8}$ of an acre)

1949–50		1950–51	
D.F.	S.S.	D.F.	S.S.
Between villages $\sum_{i=1}^{k}(v_i-1)=42$	77802	54	102164
Within villages $\sum_{i=1}^{k}\sum_{j=1}^{v_i}(2\lambda_{ij}-1)=171$	143556	184	160214

Analysis of Variance Table (Case II)

	1949–50		1950–51	
	D.F.	S.S.	D.F.	S.S.
Between pairs	$\sum_{i=1}^{k}(n_i-1)=77$	113528	92	139162
Within pairs	$\sum_{i=1}^{k} n_i=136$	107803	146	123216

Estimates of σ_b^2 and σ_w^2 derived from the two tables are given below:

	Case I		Case II	
	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_b^2$
1949–50	840	276	793	341
1950–51	871	319	844	334

and serve to indicate underestimation of the within units variance and overestimation of the between units variance by not using the correct estimation procedure.

3. If M_i (but not necessarily N) be large, both V_2 and V_1 reduce to

$$V = \frac{1}{n}\left[\sigma_b^2 + \frac{\sigma_\omega^2}{m}\right]$$

and it may be desirable to obtain values of m and n which would minimize the cost for a given variance. Under most practical situations that cost would be directly related to the distinct number v of selected primary units, which in this case is a random variable. We may therefore minimize the *expected* cost for a given variance. Let the cost function be of the form

$$C = C_2 mn + C_1 E(v),$$

where

C_1 = cost per distinct p.s.u.

C_2 = cost per sub-unit.

The optimum values of m and n are given by

$$n_0 = \lambda\sigma_b^2/(\lambda V - \sigma_\omega\sqrt{C_2})$$

$$m_0 = \sigma_\omega(\lambda V - \sigma_\omega\sqrt{C_2})/\sigma_b^2\sqrt{C_2},$$

where λ satisfies

$$-C_1 \sum_{i=1}^{N} (1 - p_i)^n \log (1 - p_i) = \frac{(\lambda V - \sigma_\omega \sqrt{C_2})^2}{\sigma_b^2}$$

or approximately λ is the real positive root of the cubic

$$C_1 - \frac{C_1 \lambda \sigma_b^2 \Sigma p_i^2}{\lambda V - \sigma_\omega \sqrt{C_2}} = \frac{(\lambda V - \sigma_\omega \sqrt{C_2})^2}{\sigma_b^2}$$

It is worth noting that

$$m_0 = \frac{\sigma_\omega}{\sigma_b} \sqrt{\frac{C_1}{C_2}} \left\{ \frac{\partial}{\partial n} E(v) \right\}_{n=n_0}$$

This combined with the equation for variance provides a simple alternative method of evaluating m_0 and n_0 by trial and error. If N also is large, the expression for m_0 reduces to the well-known result,

$$m_0 = \frac{\sigma_\omega}{\sigma_b} \sqrt{\frac{C_1}{C_2}}.$$

4. We have discussed so far the possible improvements by avoiding the repetition of sub-units in a selected primary unit. One would, however, feel that it may be still more desirable to avoid the repetition of primary units themselves. It therefore appears proper to postulate a sampling procedure which though carried out with replacement would ensure a given number of distinct units. Such a plan is described below. Though we have given its basic theory, unfortunately the results arrived at are not in a form convenient for practical application. Possibly simpler expressions could be worked out.

The new plan considered is as follows: From a population of N p.s.u.'s, the i-th unit is to be selected with probability p_i. We continue drawing units with replacement till we get a given number v distinct units. In the above process if the i-th p.s.u. is repeated i times, $m\lambda_i$ secondary units will be sub-sampled out of it, the sub-sampling being done with equal probability to each sub-unit in a primary unit. It is assumed that the number of secondary units in a primary unit is large. We shall work out certain basic results in respect of such a sampling procedure. It is to be noted that the process of sampling the p.s.u.'s is a sequential one terminating at the draw which completes v distinct units. The total number of drawings n is thus a random variable.

Lemma.—If from a population of N units with values

$$Z_1, Z_2, \ldots, Z_N$$

units are drawn with replacement until we get ν distinct units, the probability of drawing the i-th unit being p_i, and if the i-th distinct unit is repeated λ_i times in the sample, the moment generating function of $\sum_{i=1}^{\nu} \lambda_i Z_i$ is given by

$$\phi(t) = E\left\{e^{t\sum_{i=1}^{\nu}\lambda_i Z_i}\right\}$$

$$= 1 + \left\{\sum_{i=1}^{N} p_i (e^{tZ_i} - 1)\right\}\left\{\sum_{r=0}^{\nu-1} \Sigma \frac{A_r}{1-\phi_r}\right\},$$

where

$$A_r = (-1)^{n-r-1}\binom{N-r-1}{n-r-1}$$

$$\phi_r = p_{i_1} e^{tZ_{i_1}} + p_{i_2} e^{tZ_{i_2}} + \ldots + p_{i_r} e^{tZ_{i_r}}$$

and the double summation extends over all possible combinations of $\nu - 1$ or lesser number of units.

Proof.—For simplicity consider $\nu = 3$. If z_i be the last unit drawn we note that $\lambda_i = 1$.

The moment generating function of

$$Z_i + \lambda_j Z_j + \lambda_h Z_h$$

is therefore given by

$$\sum_{i,j,h} \sum_{\lambda} p_i p_j^{\lambda_j} p_h^{\lambda_h} e^{t(Z_i + \lambda_j Z_j + \lambda_h Z_h)}$$

where the summation extends over all non-zero values of λ_j and λ_h and over all possible choice of i, j, h. Putting $p_i e^{tZ_i} = \omega_i$ this reduces to

$$\sum_{i,j,h}\left[\omega_i\left\{\frac{\omega_j + \omega_h}{1 - \omega_j - \omega_h} - \frac{\omega_j}{1-\omega_j} - \frac{\omega_h}{1-\omega_h}\right\}\right]$$

$$= \left[\sum_{i=1}^{N} \omega_i - 1\right]\left\{\sum_{j,h} \frac{1}{1-\omega_j-\omega_h} - (N-2)\sum_{i} \frac{1}{1-\omega_i} + \frac{(N-1)(N-2)}{2}\right\} + 1$$

Remembering that

$$\sum_{r=0}^{\nu-1} {}^{N}C_r A_r = 1$$

the general result follows by procedure along lines exactly similar to the above.

In particular we derive the following results:

$$U \equiv E(n) = \sum_{r=0}^{\nu-1} \Sigma A_r (1 - S_r)^{-1},$$

where

$$S_r = p_{i_1} + p_{i_2} + \ldots + p_{i_r}$$
$$H \equiv E(\Sigma \lambda_i \bar{Y}_i) = UY$$

$$E(\Sigma \lambda_i \bar{Y}_i)^2 = \sum_{r=0}^{\nu-1} \Sigma \frac{K(1 - S_r) + 2YB_r}{(1 - S_r)^2} A_r$$

where

$$K = \sum_{i=1}^{N} p_i \, Y_i^2$$

and

$$B_r = p_{i_1} Y_{i_1} + p_{i_2} Y_{i_2} \ldots + p_{i_r} Y_{i_r}.$$

$$E(n^2) = \sum_{r=0}^{\nu-1} \Sigma \frac{1 + S_r}{(1 - S_r)^2}$$

$$E(n \Sigma \lambda_i Y_i) = \sum_{r=0}^{\nu-1} \Sigma \frac{Y + B_r}{(1 - S_r)^2} A_r$$

$$E(\Sigma \lambda_i^2 Y_i^2) = KU + 2 \sum_{r=0}^{\nu-1} \Sigma C_r (1 - S_r)^{-2} A_r,$$

where

$$C_r = p_{i_1}^2 Y_{i_1}^2 + p_{i_2}^2 Y_{i_2}^2 + \ldots + p_{i_r}^2 Y_{i_r}^2$$

Using these results, the variance of the ratio estimate, correct to order $\frac{1}{\nu}$ works out to

$$\frac{H^2}{U^2}\left[\frac{1}{U^2} \sum_{r=0}^{\nu-1} \Sigma \frac{1 + S_r}{(1 - S_r)^2} A_r - \frac{2}{UH} \Sigma\Sigma \frac{Y + B_r}{(1 - S_r)^2} A_r + \frac{1}{H^2} \Sigma\Sigma \frac{K(1 - S_r) + 2YB}{(1 - S_r)^2} A_r + \frac{U}{nH^2} \sigma_\omega^2\right].$$

SUMMARY

The gain in efficiency achieved by avoiding the repetition of sub-units in a two-stage sampling plan has been explicitly worked out, the percentage reduction in the within unit component of the variance being very nearly equal to the over-all sampling fraction. Expressions have been given for estimating the variance from a sample. Optimum values of m and n have been worked out which minimize the expected cost for a given variance.

A sampling procedure has been postulated which though carried out with replacement, ensures a given number of distinct primary units and an expression for variance of the estimate has been worked out.