# NON-SAMPLING ERRORS IN SURVEYS*

By P. V. Sukhatme and G. R. Seth
*Indian Council of Agricultural Research, New Delhi*

## 1. Introduction

Errors affecting the results of sample surveys can be divided in two classes:

(*a*) those arising from a fraction of the population being observed called sampling errors, and

(*b*) those arising from variable response or biases of the enumerators called non-sampling errors.

An eye estimate of the yield of a crop in a field provides an example of the source of non-sampling errors. Eye estimate is a form of measurement which cannot, in the nature of things, be unique even when the same field is observed at different times by the same enumerator. The result will depend upon the personal judgment of the enumerator and that possibly changing on different occasions, no matter how he is trained. Experience in fact indicates that a good crop is usually under-estimated and a bad one over-estimated in different degrees by men of the revenue department who are usually required to furnish eye estimates of the condition of crop in advance of the harvesting dates. Even with factual characters like the area under the crop in a field, there is found to be a marked variation in performance of the same or different enumerators. The net effect of these discrepancies on the value of the estimate is not, however, always serious, for errors very likely compensate each other although they necessarily inflate variance of the estimate. Too often, however, these errors do not cancel and the net effect is a bias which at times is much larger as compared with the sampling error than is commonly

---

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

*Note.*—Since going to the Press our attention has been drawn to a paper on 'Response Errors in Surveys' by M. H. Hansen, W. N. Hurwitz, E. S. Marks and W. P. Mauldin in June 1951 issue of *Journal of American Statistical Association* covering a somewhat similar ground. The case considered by them is covered by the general formulæ in our paper when $\sigma_{yI}$ is replaced by $\sigma_\alpha^2$ and $\sigma_y^2 - \sigma_{yI}$ by $\varphi^2 + \sigma_\delta^2 + \sigma_\epsilon^2$ the interviewers being selected at random out of a given pool of M.

believed even when all the care is bestowed in evolving suitable questionnaires and methods of measurement and in providing thorough training and supervision of field work. A striking example is furnished by the use of small plots in sample harvesting of crops.[1] Replicated samples alternatively called interpenetrating checks (I.P.C.) by Mahalanobis are used to estimate non-sampling errors. Replicated samples can be arranged in a variety of ways as follows:

(i) independent samples each enumerated by a different party of enumerators as in cases I and II of Section 3.3;

(ii) completely or partially overlapping samples as in cases III and IV of Section 3.3; and

(iii) linked samples where the selection is such that the configurations of the different samples in the sample space are not independent but each sample provides an unbiased estimate of the population character.

It is the object of this paper to suggest methods of measuring the components of non-sampling errors based upon a mathematical model which is general enough to cover the conditions commonly met in agricultural and socio-economic surveys. We will illustrate the method by giving examples from actual surveys conducted by the Indian Council of Agricultural Research.

## 2. Mathematical Model

Let $x_i$ $(i = 1, 2, \ldots l)$ denote the value of the character of the $i$-th unit in the sample of $l$ units selected randomly out of a finite or an infinite population with mean $\mu$ and variance $\sigma^2$. Also let

$$y_{ijk},\ i = 1, 2, \ldots l;$$
$$j = 1, 2, \ldots m;$$
$$k = 0, 1, 2, \ldots n_{ij}$$

be the value reported by the $j$-th enumerator on the $i$-th unit for the $k$-th occasion. It will be seen that the $j$-th enumerator has been assumed to make $n_{ij}$ observations on the $i$-th unit and the number of enumerators participating in the survey is $m$. We assume that $y_{ijk}$ is made up of four components as follows:

$$y_{ijk} = x_i + \alpha_j + \delta_{ij} + \epsilon_{ijk}, \tag{1}$$

where

$x_i$ denotes the intrinsic value of the $i$-th unit,

$a_j$ represents the bias of the $j$-th enumerator in repeated observations on all the units in the population,

$a_j + \delta_{ij}$ represents the bias of the $j$-th enumerator in repeated observations on the $i$-th unit,

$a_j + \delta_{ij} + \epsilon_{ijk}$ represents the total deviation from $x_i$ when $j$-th enumerator reports on the $i$-th unit for the $k$-th occasion.

It follows that

$$\left.\begin{aligned} E(\epsilon_{ijk} \mid i, j) &= 0 \\ E(\delta_{ij} \mid j) &= 0 \end{aligned}\right\}, \tag{2}$$

where $E(X)$ denotes the expectation of $X$.

It may be mentioned that $\delta_{ij}$ represents the interaction of the $j$-th enumerator with the $i$-th unit and its value can be zero or vary with the size of the unit in accordance with some known law, linear for example, or unknown. When $\delta_{ij}$ is zero, equation (1) reduces to

$$y_{ijk} = x_i + a_j + \epsilon_{ijk}. \tag{3}$$

For the linear law, $y_{ijk}$ will reduce to

$$y_{ijk} = x_i (1 + r_j) + a_j + \epsilon_{ijk}. \tag{4}$$

In this paper we will confine our attention to model given by (1). Model given by (3) is a special case of (1). The consequences of model given by (4) have also been worked out but space does not permit their discussion.

## 3. Estimates of the Components of the Bias with Errors

3.1. To start with, we will consider the population consisting of one stratum only.

Let

$\bar{y}_{ij}$ represent the mean of the $n_{ij}$ observations made on the $i$-th unit by the $j$-th enumerator,

$\bar{y}_{.j.}$ the mean of all the observations, $n_{.j} = \sum_{i=1}^{l} n_{ij}$ made by the $j$-th enumerator,

$\bar{y}_{i..}$ the mean of all the observations $n_{i.} = \sum_{j=1}^{m} n_{ij}$ made on the $i$-th unit,

$\bar{y}_{...}$ the mean of all observations $n = \sum_{j} \sum_{i} n_{ij}$ made on all the $l$ units in the sample.

With a similar notation for the means of $\epsilon_{ijk}$'s, we have

$$\left.\begin{aligned} y_{ij.} &= x_i + \alpha_j + \delta_{ij} + \epsilon_{ij.} \\ y_{.j.} &= \frac{\sum_{i=1}^{l} x_i n_{ij}}{n_{.j}} + \alpha_j + \frac{\sum_{i=1}^{l} \delta_{ij} n_{ij}}{n_{.j}} + \epsilon_{.j.} \\ y_{i..} &= x_i + \frac{\sum_{j=1}^{m} \alpha_j n_{ij}}{n_{i.}} + \frac{\sum_{j=1}^{m} \delta_{ij} n_{ij}}{n_{i.}} + \epsilon_{i..} \\ y_{...} &= \frac{\sum_{i=1}^{l} x_i n_{i.}}{n} + \frac{\sum_{j=1}^{m} \alpha_j n_{.j}}{n} + \frac{\sum_{i}\sum_{j} \delta_{ij} n_{ij}}{n} + \epsilon_{...} \end{aligned}\right\} \quad (5)$$

Two situations have to be considered in the measurement of the bias according as (*a*) the $m$ enumerators are fixed or (*b*) randomly selected out of a larger population $M$. We will first consider case (*a*). The different enumerators may come from different agencies as for instance agricultural and the revenue agencies in case of crop surveys or $m$ different schools using different methods of enumeration. The expectations and the variances of equation (5) are given by

$$\left.\begin{aligned} E(y_{ij.} \mid j) &= \mu + \alpha_j \\ E(y_{.j.} \mid j) &= \mu + \alpha_j \\ E(y_{i..} \mid j = 1, 2, \ldots, m) &= \mu + \sum_{j=1}^{m} \alpha_j n_{ij}/n_{i.} \\ E(y_{...} \mid j = 1, 2, \ldots, m) &= \mu + \sum_{j=1}^{m} \alpha_j n_{.j}/n \end{aligned}\right\} \quad (6)$$

and

$$\left.\begin{aligned} V(y_{ij.} \mid j) &= \sigma^2 + \sigma^2_{\delta_j} + \frac{\sigma^2_{\epsilon}}{n_{ij}} \\ V(y_{.j.} \mid j) &= \sigma^2 \sum_{i=1}^{l} n_{ij}^2/n_{.j}^2 + \sum_{i=1}^{l} n_{ij}^2 \sigma^2_{\delta_j}/n_{.j}^2 + \frac{\sigma^2_{\epsilon}}{n_{.j}} \\ V(y_{i..} \mid j = 1, 2, \ldots, m) &= \sigma^2 + \frac{\sum_{j=1}^{m} n_{ij}^2 \sigma^2_{\delta_j}}{n_{i.}^2} + \frac{\sigma^2_{\epsilon}}{n_{i.}} \\ V(y_{...} \mid j = 1, 2, \ldots, m) &= \frac{\sigma^2 \sum_{i=1}^{l} n_{i.}^2}{n^2} + \frac{\sum_{i}\sum_{j} n_{ij}^2 \sigma^2_{\delta_j}}{n^2} + \frac{\sigma^2_{\epsilon}}{n} \end{aligned}\right\} \quad (7)$$

neglecting the term $-\frac{\sigma^2}{N}$ in case the population is finite consisting of $N$ units.

where $\sigma^2$, as already stated, is the variance of $x_i$ given by

$$\sigma^2 = E(x_i - \mu)^2 \text{ or } \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N-1}$$

and

$$\sigma_{\delta_i}^2 = E[\{\delta_{ij} - E(\delta_{ij} \mid j)\}^2 \mid j],$$

$$\sigma_\epsilon^2 = E[\{\epsilon_{ijk} - E(\epsilon_{ijk} \mid i, j)\}^2 \mid i, j],$$

$\sigma_\epsilon^2$ being assumed constant for all $i$ and $j$. Before evaluating the biases or their errors, we offer a few remarks on equations (6) and (7). If $a_j$ is the same for all enumerators, then the bias in the average cannot be removed by employing more than one enumerator. On the other hand if $a_j$ is variable so that $\frac{\Sigma a_j n_{.j}}{n}$ is negligible as the number of enumerators is increased adequately, and though the estimate, is free from bias, its sampling error is not entirely due to the inherent variability $\sigma^2$ but is inflated by the variability in response due to terms involving $\sigma_{\delta_j}^2$ and $\sigma_\epsilon^2$. This emphasises the fact that it is not sufficient to ensure that $a_j$'s are variable and cancel each other on the average but also to see that the effect of the variability in response on the sampling variation is reduced to the minimum by taking maximum care in planning surveys.

3.2. For evaluating $\sigma_\epsilon^2$, $\sigma_{\delta_j}^2$ and $\sigma^2$, we give below the expectations of the different mean squares in the sample:

$$E[S_\omega^2 = \sum_{i,j,k} (y_{ijk} - y_{ij.})^2/(n-c)] = \sigma_\epsilon^2, \tag{8}$$

where $c$ represents the number of non-zero $n_{ij}$'s ($i = 1, 2, \ldots, l$; $j = 1, 2, \ldots, m$),

$$E\left[S_a^2 = \frac{\sum\sum_{j} n_{ij}(y_{ij.} - y_{...})^2}{c-1}\right]$$

$$= \frac{n - \frac{\sum_i n_{i.}^2}{n}}{c-1}\sigma^2 + \frac{\sum_j n_{.j}\left(a_j - \frac{\sum_j a_j n_{.j}}{n}\right)^2}{c-1}$$

$$+ \frac{\sum_j n_{.j}\sigma_{\delta_j}^2 - \frac{\sum_{i,j} n_{ij}^2 \sigma_{\delta_j}^2}{n}}{c-1} + \sigma_\epsilon^2. \tag{9}$$

$$E\left[S_e^2 = \frac{1}{m-1}\sum_{j=1}^{m} n_{.j}\,(y_{.j.} - y_{...})^2\right]$$

$$= \frac{\sigma^2}{m-1}\left[\sum_{i,j}\frac{n_{ij}^2}{n_{.j}} - \frac{\sum_i n_{i.}^2}{n}\right] + \frac{1}{m-1}\sum_j n_{.j}\left(a_j - \frac{\sum_j a_j n_{.j}}{n}\right)^2$$

$$+ \frac{\sum_{i,j}\frac{n_{ij}^2\sigma_{\delta j}^2}{n_{.j}} - \sum_{i,j}\frac{n_{ij}^2\sigma_{\delta j}^2}{n}}{m-1} + \sigma_\epsilon^2 \qquad (10)$$

$$E\left[S_0^2 = \frac{\sum_i n_{i.}\,(y_{i..} - y_{...})^2}{l-1}\right] = \frac{1}{l-1}\left[n - \frac{\sum_i n_{i.}^2}{n}\right]\sigma^2$$

$$+ \frac{\sum_{i=1}^{l} n_{i.}\left(\frac{\sum_j a_j n_{ij}}{n_{i.}} - \frac{\sum_j a_j n_{.j}}{n}\right)^2}{l-1} + \frac{\sum_{i,j}\frac{n_{ij}^2}{n_{i.}}\sigma_{\delta j}^2 - \frac{\sum_{i,j} n_{ij}^2\sigma_{\delta j}^2}{n}}{l-1}$$

$$+ \sigma_\epsilon^2 \qquad (11)$$

$$E\left[S_{j0}^2 = \frac{\sum_i (y_{ij.} - y_{.j.})^2\, n_{ij}}{c_j - 1}\,\middle|\, j\right] = \frac{n_{.j} - \sum_i \frac{n_{ij}^2}{n_{.j}}}{c_j - 1}\,\sigma^2$$

$$+ \frac{1}{c_j - 1}\left[n_{.j}\,\sigma_{\delta j}^2 - \sum_i \frac{n_{ij}^2}{n_{.j}}\,\sigma_{\delta j}^2\right] + \sigma_\epsilon^2 \qquad (12)$$

$$E\left[(S_{c0}^2) = S_{cc}^2 = \frac{\sum_{i,j} (y_{ij.} - y_{.j.})^2\, n_{ij}}{c-m}\right]_{l=1,\,2,\,\ldots,\,m}$$

$$\frac{n - \sum_{i,j}\frac{n_{ij}^2}{n_{.j}}}{c-m}\,\sigma^2 + \frac{1}{c-m}\left[\sum_j n_{.j}\,\sigma_{\delta j}^2 - \sum_{i,j}\frac{n_{ij}^2\,\sigma_{\delta j}^2}{n_{.j}}\right] + \sigma_\epsilon^2 \qquad (13)$$

where $S_w^2$, $S_c^2$, $S_e^2$, $S_0^2$, $S_{j0}^2$ and $S_{e0}^2$ represent mean squares within cells (a cell being determined by $i$, $j$), between cells, between enumerators, between units, between units within the $j$-th enumerator and between units within enumerators respectively. In case $\sigma_{\delta j}^2$ is the same for all $j$, we will replace $\sigma_{\delta j}^2$ by $\sigma_\delta^2$ in equations (9), (11) and (12) will then not be needed, for (13) will be sufficient for our purpose. Wherever $\sigma_\alpha^2$ and $\sigma_{\delta_i}^2$ are separately estimated, appropriate linear functions of the mean squares also provide tests of significance for the departures from zero.

3.3. So far we have not imposed any restriction on $n_{ij}$'s but in actual sample survey work $n_{ij}$'s follow a simple and systematic pattern both for convenience of field work and analysis. We now give a few designs which may be used in sample surveys.

*Case I.*—$n_{i\cdot} = 1$, *i.e.*, a unit is observed once only, randomly allotted to one observer or the other with the restriction that

$$n_{\cdot j} = \frac{n}{m}$$

$S^2_W$ does not exist and we cannot separately estimate $\sigma^2$, $\sigma_\epsilon^2$, and $\sigma_{\delta_j}^2$ but only the sum of the three together, *viz.*,

$$\sigma^2 + \sigma_\epsilon^2 + \sigma_{\delta_j}^2$$

The analysis of variance takes the form

| Source | D.F. | M.S. | Is the estimate of |
|---|---|---|---|
| Between enumerators | $m-1$ | $S_c^2$ | $\sigma^2 + \frac{n}{m}\sigma_\alpha^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |
| Within enumerators .. | $n-m$ | $S_{e0}^2$ | $\sigma^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |
| Total .. | $n-1$ | $S_0^2$ | $\sigma^2 + \frac{n(m-1)}{m(l-1)}\sigma_\alpha^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |

In the above $\sigma_\alpha^2$, $\bar{\sigma}_\delta^2$ are given by

$$\sum_{j=1}^{m} \frac{\left(a_j - \frac{\Sigma a_j}{m}\right)^2}{m-1} \text{ and } \frac{\sum_{j=1}^{m} \sigma_{\delta_j}^2}{m}.$$

From the table of analysis of variance we obtain

$$\frac{m}{n}(S_c^2 - S_{e0}^2)$$

as the unbiased estimate of $\sigma_\alpha^2$. $S_c^2/S_{e0}^2$ also provides a test for the hypothesis that $\sigma_\alpha^2 = 0$. Sampling variance of $y_{\cdot\cdot\cdot}$, the general mean is estimated by $S_{e0}^2/n$. The total error which we define as the mean square error of the estimate of $y\ldots$ is given by

$$-\frac{\sigma^2}{N} + \frac{\sigma^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2}{n} + \bar{\alpha}^2, \text{ where } \bar{\alpha} = \sum_{l=1}^{m} a_j/m.$$

In certain cases only one enumerator, say the $j$-th, will be employed to enumerate the sample of $n$ units, the total error then is given by

$$-\frac{\sigma^2}{N} + \frac{\sigma^2 + \sigma_{\delta_j}^2 + \sigma_\epsilon^2}{n} + a_j^2.$$

The efficiency of the design compared to the case when the $j$-th enumerator is used to enumerate all the $n$ observations as measured by the

ratio of the variances of the estimates excluding considerations of the cost, is given by

$$\frac{\frac{\sigma^2+\sigma_{\delta j}^2+\sigma_\epsilon^2}{n}+a_j^2}{-\frac{\sigma^2}{N}+\frac{\sigma^2+\bar{\sigma}_\delta^2+\sigma_\epsilon^2}{n}+\bar{a}^2}$$

which can be less or greater than one depending on the $j$-th enumerator selected. On the average, however, the efficiency is given by

$$\frac{\frac{\sigma^2+\bar{\sigma}_\delta^2+\sigma_\epsilon^2}{n}+\sum_{j=1}^{m} a_j^2/m}{-\frac{\sigma^2}{N}+\frac{\sigma^2+\bar{a}_\delta^2+\sigma_\epsilon^2}{n}+\bar{a}^2}=1+\frac{\frac{m-1}{m}\sigma_a^2}{\frac{\sigma^2+\bar{\sigma}_\delta^2+\sigma_\epsilon^2}{n}+\bar{a}^2-\frac{\sigma^2}{N}}$$

which is greater than one. The gain in efficiency is estimated by

$$(m-1)\frac{(S_e^2-S_{e0}^2)}{S_{e0}^2}$$

if we assume that

$$E\left(\frac{S_e^2}{S_{e0}^2}\right)=\frac{E(S_e^2)}{E(S_{e0}^2)}$$

and

$$\bar{a}^2-\frac{\sigma^2}{N} \text{ is negligibile.}$$

This formula when $\bar{a}$ is not equal to zero, overestimates the efficiency of this design over the one in which only one enumerator is employed for the same number of $n$ units in both the designs. In subsequent cases also the gain in efficiency for the replicated samples averaged over all the enumerators is of the same order as in the case where one enumerator only is employed instead of $m$. Usually, however, the number of enumerators cannot be reduced below $m$ as a smaller number cannot complete the work during the period of the survey. In the latter case, if the enumerators are not allotted independent samples but given neighbouring units, there is not only no loss in efficiency compared with the replicated samples but there is reduction in the cost by this systematic allotment. The disadvantage, however, is that we cannot estimate $\sigma_a^2$ or the extent of differential bias.

*Case II.*—

$$n_{i.}=p \text{ and } n_{ij}=0 \text{ or } p$$

*i.e.*, a unit is observed $p$ times and by the same enumerator. It is also stipulated that an equal number of units are allotted randomly

to each of the enumerators. In this case $n = pl$ and $n_{ij} = \frac{pl}{m}$ and the analysis of variance takes the form

| Source | D.F. | M.S. | Is the estimate of |
|---|---|---|---|
| Between enumerators | $m-1$ | $S_e^2$ | $p\sigma^2 + \frac{n}{m}\sigma_\alpha^2 + p\bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |
| Between units within enumerators | $l-m$ | $S_{e0}^2$ | $p\sigma^2 + p\bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |
| Within units | $(p-1)\,l$ | $S_w^2$ | $\sigma_\epsilon^2$ |

The advantage of this case over I is that $\sigma_\epsilon^2$ can be estimated separately, though $\sigma^2$ and $\bar{\sigma}_\delta^2$ are still separately not estimable. The estimate of the gain in efficiency is given by

$$(m-1)\,\frac{(S_e^2 - S_{e0}^2)}{S_{e0}^2}$$

*Case III.*—

$$n_{i.} = p,\ n_{ij} = 0 \text{ or } 1,$$

*i.e.*, a unit is observed once each by $p$ enumerators. It is assumed that the units are allotted at random to the enumerators in such a way that each enumerator observes the same number of units and the number of times a pair of enumerators observe a common unit is the same for all pairs and is given by $\lambda$. The analysis of variance table in this case is as follows:

| Source | D.F. | M.S. | Is the estimate of |
|---|---|---|---|
| Between enumerators | $m-1$ | $S_0^2$ | $\sigma^2\frac{(m-p)}{m-1} + \frac{n}{m}\sigma_\alpha^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |
| Between units within enumerators | $n-m$ | $S_{e0}^2$ | $\sigma^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2$ |

This has got to be supplemented by additional information such as mean square between units in order to be able to estimate $\sigma^2$ and $\sigma_\alpha^2$ separately. We see that $\bar{\sigma}_\delta^2$ and $\sigma_\epsilon^2$ cannot be individually estimated. The expectation of the mean square between units will be given by

$$p\sigma^2 + \frac{(n - m\lambda)\,(m-1)}{pm\,(l-1)}\,\sigma_\alpha^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2,$$

where $\lambda$ represents the number of times a pair of enumerators observe the same unit. If compared to the design in which the number of units remain the same, the number of observations remaining the same

but to be enumerated by one enumerator, the gain in efficiency of this design is given by

$$\frac{\frac{m-1}{m}\sigma_\alpha^2 + \frac{p-1}{n}\bar{\sigma}_\delta^2}{\frac{\sigma^2}{l} + \frac{\bar{\sigma}_\delta^2 + \sigma_\epsilon^2}{n} + \bar{\alpha}^2 - \frac{\sigma^2}{N}}$$

The expression for estimating the gain in efficiency in case (I) and (II) will give an upper-limit for the efficiency in this case.

*Case IV.*—

$$n_{i\cdot} = 1 \text{ or } 2 \text{ and } n_{ij} = 0, 1 \text{ or } 2,$$

*i.e.*, some of the units are observed once only and the rest twice. In case a unit is observed twice, it may be done by the same person or two different persons. We allot the observations randomly under the restrictions (1) the number of units observed only once is the same for each enumerator, (2) number of units observed twice by the same enumerator is the same for all enumerators, (3) number of units observed twice but by two enumerators is constant for each pair of enumerators. From this design it follows that

$$l = ml_1 + ml_2 + \frac{m(m-1)}{2}\lambda$$

$$n = ml_1 + 2ml_2 + m(m-1)\lambda$$

$$n_{\cdot j} = l_1 + 2l_2 + (m-1)\lambda = \frac{n}{m}$$

where

$l_1$ denotes the number of units (observed once only) observed by each enumerator,

$l_2$ denotes the number of units (observed twice) observed by each enumerator,

$\lambda$ denotes the number of units observed by each pair of enumerators.

The advantage of this design over the previous ones is that it admits the possibility of estimating separately the components $\sigma^2$, $\sigma_\alpha^2$, $\bar{\sigma}_\delta^2$ and $\sigma_\epsilon^2$. On the other hand the design is not orthogonal and there are several alternate estimates of $\sigma^2$, $\sigma_\alpha^2$, $\bar{\sigma}_\delta^2$ and $\sigma_\epsilon^2$ which can be suggested. We give a set of mean squares with their expectations from which we can separately estimate these quantities. They are as follows

$$
\left.
\begin{aligned}
ES_w^2 &= \sigma_\epsilon^2 \\
ES_c^2 &= \frac{\sigma^2\left(n-3+\frac{2l}{n}\right)}{c-1}\frac{}{} + \frac{n(m-1)}{m(c-1)}\sigma_\alpha^2 + \frac{\bar{\sigma}_\delta^2\left(n-3+\frac{2c}{n}\right)}{c-1} \\
&\quad + \sigma_\epsilon^2 \\
ES_e^2 &= \frac{\sigma^2}{m-1}\left[\frac{3n-2c}{n}\,m-3+\frac{2l}{n}\right] + \frac{n}{m}\sigma_\alpha^2 + \frac{3n-2c}{n}\bar{\sigma}_\delta^2 \\
&\quad + \sigma_\epsilon^2 \\
ES_{cw}^2 &= \frac{1}{c-m}\left[n-\frac{m}{n}(3n-2c)\right]\left[\sigma^2+\bar{\sigma}_\delta^2\right] + \sigma_\epsilon^2.
\end{aligned}
\right\} \qquad (7)
$$

Total error of the general mean in this case is given by

$$\frac{\sigma^2}{n^2}(3n-2l) + \frac{3n-2c}{n^2}\bar{\sigma}_\delta^2 + \frac{\sigma_\epsilon^2}{n} + \bar{\alpha}^2 - \frac{\sigma^2}{N}$$

whereas the total error, when only one enumerator is employed, to make $n$ observations, will depend upon whether they correspond to $n$ units or less. For the simplest case when they represent $n$ units, the average total error is given by

$$\frac{\sigma^2+\bar{\sigma}_\delta^2+\sigma_\epsilon^2}{n} + \frac{\sum_{j=1}^{m}\alpha_j^2}{m} - \frac{\sigma^2}{N}.$$

Compared to this case, the case where the enumerator enumerates $ml_1$ units once and $ml_2 + \frac{m(m-1)}{2}\,l_3$ units twice, the total error is given by

$$\frac{ml_1+4ml_2+2m(m-1)\,l_3}{n^2}(\sigma^2+\bar{\sigma}_\delta^2) + \frac{\sigma_\epsilon^2}{n} + \frac{\sum_{j=1}^{m}\alpha_j^2}{m} - \frac{\sigma^2}{N}$$

which is the same as

$$-\frac{\sigma^2}{N} + \frac{3n-2l}{n^2}(\sigma^2+\bar{\sigma}_\delta^2) + \frac{\sigma_\epsilon^2}{n} + \sum_{j=1}^{m}\alpha_j^2/m.$$

3.4. In case $m$ enumerators are selected *ad hoc* for conducting the surveys and can be regarded a random sample out of a population of $M$ enumerators, the expectations of different means and variances of the means and expectations of different mean squares in 3.1 and 3.2 will need some modification. Expected values of the means will be given by

$$
\begin{aligned}
E(y_{i..} \mid m) &= \mu + \alpha \\
E(y_{...} \mid m) &= \mu + \alpha
\end{aligned}
\qquad (14)
$$

where

$$\alpha = \sum_{j=1}^{M} \alpha_j / M$$

For the variance of the means, we shall have to add

$$\sigma_\alpha^2 \left(1 - \frac{1}{M}\right), \ \sigma_\alpha^2 \left(1 - \frac{1}{M}\right),$$

$$\frac{\sum_j n_{ij}^2}{n_{i.}^2} \sigma_\alpha^2 - \frac{\sigma_\alpha^2}{M} \text{ and } \frac{\sum_j n_{.j}^2}{n^2} \sigma_\alpha^2 - \frac{\sigma^2 \alpha}{M}$$

to the equations in (7) where

$$\sigma_\alpha^2 = \sum_{j=1}^{M} (\alpha_j - \alpha)^2 / M - 1.$$

In the mean squares

$$S_e^2, \ s_v^2 \text{ and } S_0^2$$

we replace the terms containing $\alpha$'s by

$$\sigma_\alpha^2 (n - \sum_j n_{.j}^2/n)/c - 1, \ \sigma_\alpha^2 (n - \sum_j n_{.j}^2/n)/m - 1 \text{ and}$$

$$\frac{\sigma_\alpha^2 \left[ \sum_{i,j} \frac{n_{ij}^2}{n_{i.}} - \frac{\sum n_{.j}^2}{n} \right]}{l - 1}$$

respectively.

It may be remarked that in comparing the efficiencies of the design with $m$ enumerators randomly selected out of $M$ with alternative designs involving one enumerator, the latter would also be considered to be selected randomly out of $M$. In consequence, there is no bias in the means and the efficiency will be given by the ratio of the respective variances.

## 4. Estimates of the Components of Bias

L *Strata.*—Let $j^s$ ($j = 1, 2, \ldots, m$ and $s = 1, 2, \ldots, L$) denote the $j$-th enumerator in the $s$-th stratum. $1^s, 2^s, \ldots, m^s$ ($s = 1, 2, \ldots, L$) form $m$ distinct parties of $L$ enumerators each. In particular these different parties may be identified with different agencies as mentioned earlier or they may not be so identified in which case the division into $m$ parties becomes artificial and any one of the $m^L$ possible combinations may be termed a party.

4.1. *The Case of Fixed Enumerators for each Stratum.*—Let

$x_i^s$ $(i = 1, 2, \ldots, l_s)$ be the $l_s$ units randomly selected for enumeration by the different enumerators in the $s$-th stratum, $(s = 1, 2, \ldots, L)$

$n_{ij}^s$ represent the number of times $x_i^s$ is observed by the enumerator $j^s$ in the $s$-th stratum,

$y_{ijk}^s$ $(k = 1, 2, \ldots, n_{ij}^s)$ be the value reported by the $j$-th enumerator on $x_i^s$ on the $k$-th occasion.

We assume as before

$$y_{ijk}^s = x_i^s + \alpha_j^s + \delta_{ij}^s + \epsilon_{ijk}^s$$

where the different components on the right have the same meanings given in Section 2. Using the notation in the previous sections for the means and the variances except for the introduction of the letter $s$, to indicate the $s$-th stratum, we have, under the assumption that $\frac{n^s}{n} = \frac{N^s}{N} = p_s$, where $N^s$ and $N$ represent the units in the $s$-th stratum and all strata put together.

$$\left.\begin{aligned}
&E(y_{.j.}^s \mid j^s) = \mu_s + \alpha_j^s \\
&E(y_{...}^s \mid 1^s, 2^s, \ldots j^s \ldots m^s) = \mu_s + \frac{\sum_{j=1}^{m} \alpha_j^s n_{.j}^s}{n_{..}^s} \\
&E(y_{.j.} \mid j^1, j^2, \ldots, j^L) = \frac{\Sigma \mu_s n_{.j}^s}{n_{.j}} + \frac{\Sigma \alpha_j^s n_{.j}^s}{n_{.j}} \\
&E\begin{pmatrix} j = 1, 2, \ldots, m \\ (y_{...} \mid j^s;\ s = 1, 2, \ldots, L) \end{pmatrix} \\
&\qquad = \frac{\Sigma \mu_s n_{..}^s}{n} + \frac{\Sigma\Sigma \alpha_j^s n_{.j}^s}{n}
\end{aligned}\right\} \quad (15)$$

and

$$\left.\begin{aligned}
V(y_{.j.}{}^{s} \mid j^{s}) &= \frac{\sum_{i=1}^{l_s} (n_{ij}{}^{s})^2\, \sigma_s{}^2}{(n_{.j}{}^{s})^2} + \frac{\sum_{i=1}^{l_s} (n_{ij}{}^{s})^2}{(n_{.j}{}^{s})^2}\, \sigma_{\delta_{js}}{}^2 + \frac{\sigma_\epsilon{}^2}{n_{.j}{}^{s}} \\
V(y_{...}{}^{s} \mid 1^{s}, 2^{s}, \ldots, m^{s}) &= \frac{\sum_{i=1}^{l_s} (n_{i.}{}^{s})^2\, \sigma_s{}^2}{(n_{..}{}^{s})^2} + \frac{\sum_{i,j} (n_{ij}{}^{s})^2}{(n_{..}{}^{s})^2} + \frac{\sigma_\epsilon{}^2}{n_{..}{}^{s}} \\
V(y_{.i.} \mid j^{1}, j^{2}, \ldots, j^{L}) &= \frac{\sum_{s,\,i} (n_{ij}{}^{s})^2\, \sigma_s{}^2}{(n_{.j})^2} + \frac{\sum_{s,\,i} (n_{ij}{}^{s})^2\, \sigma_{\delta_{js}}{}^2}{(n_{.j})^2} + \frac{\sigma_\epsilon{}^2}{n_{.j}} \\
&\quad - \frac{\sum_{s=1}^{L} p_s\, \sigma_s{}^2}{N} \\
V\left(y_{...} \,\middle|\, \begin{matrix} j^{s},\ j = 1, 2, \ldots, m \\ s = 1, 2, \ldots, L \end{matrix}\right) &= \frac{\sum_{s,\,i} (n_{i.}{}^{s})^2\, \sigma_s{}^2}{n^2} + \frac{\sum_{s,\,j,\,i} (n_{ij}{}^{s})^2\, \sigma_{\delta_{js}}{}^2}{n^2} \\
&\quad + \frac{\sigma_\epsilon{}^2}{n} - \frac{\sum_{s=1}^{L} p_s\, \sigma_s{}^2}{N}
\end{aligned}\right\} \qquad (16)$$

The different mean squares $S_w{}^2$, $S_{sc}{}^2$, $S_{s0}{}^2$, $S_{se}{}^2$ and $S_{(se)c}{}^2$ will be obtained from equations (8)–(13) by summing the corresponding "sum of squares" over all strata divided by the total number of degrees of freedom so obtained. The mean squares between strata and parties have the expectations given by

$$\begin{aligned}
E\left[S_s{}^2 = \frac{\sum_{s=1}^{L} n_{..}{}^{s} (y_{i..}{}^{s} - y_{...})^2}{L-1}\right] &= \sum_{s,\,j}\left[\frac{(n_{i.}{}^{s})^2}{n_{..}{}^{s}} - \frac{(n_{i.}{}^{s})^2}{n}\right] \sigma_s{}^2/L - 1 \\
&+ \sum_{s} n_{..}{}^{s}\left[\sum_{j} \frac{(\mu_s + \alpha_j{}^{s})\, n_{.j}{}^{s}}{n_{..}{}^{s}} - \sum_{s,\,j} \frac{(\mu_s + \alpha_j{}^{s})\, n_{.j}{}^{s}}{n}\right]^2 \Big/ L - 1 + \ldots \\
&+ \sum_{s,\,j,\,i}\left[\frac{(n_{ij}{}^{s})^2}{n_{..}{}^{s}} - \frac{(n_{ij}{}^{s})^2}{n}\right] \sigma_{\delta_{js}}{}^2/L - 1 + \sigma_\epsilon{}^2 \\
&- \sum_{s=1}^{L} \frac{\sigma_s{}^2}{N_s\,(L-1)}\left(n_{..}{}^{s} - \frac{(n_{..}{}^{s})^2}{n}\right)
\end{aligned} \qquad (17)$$

$$E\left[S_p^2 = \frac{\sum_{j=1}^{m} n_{.j}\,(y_{.j} - y_{..})^2}{m-1}\right] = \frac{\sum_{s,j,i} n_{.j}\left(\frac{n_{ij}^{s}}{n_{.j}} - \frac{n_{i.}^{s}}{n}\right)^2 \sigma_s^2}{m-1} +$$

$$+ \sum_{j} n_{.j}\left(\frac{\sum_{s}(\alpha_j^s + \mu_s)\, n_{.j}^{s}}{n_{.j}} - \frac{\sum_{s,j}(\alpha_j^s + \mu_s)\, n_{.j}^{s}}{n}\right]^2 \Big/ m-1$$

$$+ \sum_{s,j,i} n_{.j}\left(\frac{(n_{ij}^{s})^2}{n_{.j}} - \frac{(n_{ij}^{s})^2}{n}\right]^2 \sigma_{\delta_{js}}^2/m-1 + \sigma_\epsilon^2. \qquad (18)$$

In case $\frac{n_{.j}^{s}}{n^{s}}$ is independent of $j$, then the term containing $\alpha$'s in (18) is independent of $\mu$'s and reduces to

$$\sum_{j} n_{.j}\left(\frac{\sum_{s}\alpha_j^s n_{..}^{s}}{n} - \frac{\sum_{s,j}\alpha_j^s n_{.j}^{s}}{n}\right)^2 \Big/ m-1 \qquad (19)$$

4.2. The treatment of the special cases corresponding to the four designs studied in Section 3·3 for one stratum follows identical lines. We now take up the question of efficiency of various alternative designs. We assume that $\sigma_s^2 = \sigma^2$ and $\sigma_{\delta_{js}}^2 = \sigma_{\delta_j}^2$.

*Special Cases*

*Case I.*—

$$n_{i.}^{s} = 1$$

$$n_{ij}^{s} = 0 \text{ or } 1$$

$$\frac{n_{.j}^{s}}{n_{..}^{s}} = 1/m$$

and

$$\frac{n_{..}^{s}}{n} = \frac{N^{s}*}{N}$$

The total error of the general mean when $m$ parties work is given by

$$-\frac{\sigma^2}{N} + \frac{\sigma^2}{n} + \frac{\bar{\sigma}_\delta^2}{n} + \frac{\sigma_\epsilon^2}{n} + \left(\sum_{s,j}\frac{\alpha_j^s n_{.j}^{s}}{n}\right)^2$$

while the total error in case some one party does the work is, on the average, given by

---

* When it is not so, the strata means have to be weighted by $p_s = \frac{N^s}{N}$ to get the unbiased estimates where there are no biases.

$$- \frac{\sigma_\alpha^2}{N} + \frac{\sigma^2}{n} + \frac{\bar{\sigma}_\delta^2}{n} + \frac{\sigma_\epsilon^2}{n} + \sum_{j=1}^{m} \left( \frac{\sum_{s=1}^{L} a_j^s n_{.j}^s}{n_{.j}} \right)^2 \Big/ m$$

The gain in efficiency of $m$ parties compared to the case in which one party is employed has the expression

$$\frac{\frac{1}{m} \sum_{j=1}^{m} n_{.j} \left[ \frac{\sum_s a_j^s n_{.j}^s}{n_{.j}} - \frac{\sum_s \sum_j a_j^s n_{.j}^s}{n} \right]^2}{\frac{\sigma^2}{n} - \frac{\sigma_\alpha^2}{N} + \frac{\sigma_\epsilon^2}{n} + \frac{\bar{\sigma}_\delta^2}{n} + \frac{(\sum a_j^s n_{.j}^s)^2}{n^2}}$$

which is estimated by

$$\frac{(m-1)\,[S_p^2 - S_{e0}^2]}{S_{e0}^2}$$

on the assumption that $\left(\bar{a}^2 - \frac{\sigma^2}{N}\right)$ is negligible,

where

$$\bar{a} = \sum_{s,j} \cdot \frac{a_j^s n_{.j}^s}{n}.$$

It may be noticed that the expression is the same as that for the corresponding case of one stratum if we replace the enumerator by a party. Cases II, III and IV yield exactly the same expressions for efficiency as that given in Section 3.3 for a single stratum when $S_e^2$ is replaced by $S_p^2$.

4.3. For the case when the members of the $j$-th party are randomly selected from the population of $M_j$ enumerators, with mean $\alpha_j$ and variance

$$\sigma_{j\alpha}^2 = \sum_{j=1}^{M_j} \left( a_j^s - \frac{\sum_{s=1}^{M_j} a_j^s}{M_j} \right)^2 \Big/ M_j - 1$$

The expectations of the means take the form

$$E(y_{.j}^s) = \mu_s + \alpha_j$$

$$E(y_{.j}) = \frac{\sum_{s=1}^{L} \mu_s n_{.j}^s}{n_{.j}} + \alpha_j$$

$$E(y_{..}^{s}) = \mu_s + \frac{\sum_{j=1}^{m} \alpha_j n_{.j}^{s}}{n_{..}^{s}}$$

$$E(y_{...}) = \frac{\sum_{s=1}^{L} \mu_s n_{..}^{s}}{n} + \frac{\sum_{j=1}^{m} \alpha_j n_{.j}}{n}$$

The variances of the means are the same as those in (16) except for the addition of

$$\sigma_{j\alpha}^2 \left(1 - \frac{1}{M_j}\right), \frac{\sum_{j=1}^{m} \sigma_{j\alpha}^2 \left(1 - \frac{1}{M_j}\right) (n_{.j}^{s})^2}{(n_{..}^{s})^2}, \sigma_{j\alpha}^2 \left[\frac{\sum_{s=1}^{L} (n_{.j}^{s})^2}{(n_{.j})^2} - \frac{1}{M_j}\right]$$

and

$$\sum_{j=1}^{m} \sigma_{j\alpha}^2 \left[\frac{\sum_{s=1}^{L} (n_{.j}^{s})^2}{n^2} - \frac{1}{M_j} \frac{(n_{.j})^2}{n^2}\right]$$

to the different equations. As for the expected values of different mean squares, they will be obtained by changing $\alpha_j^s$ to $\alpha_j$ and adding an extra term to each of the different mean squares for the case of fixed enumerators. The extra terms are given as follows:

$S_w^2$ .. No addition

$$S_{sc}^2 \quad .. \quad \sum_{j=1}^{m} \left(n_{.j} - \sum_{s} \frac{(n_{.j}^{s})^2}{n_{..}^{s}}\right)\left(1 - \frac{1}{M_j}\right) \sigma_{j\alpha}^2/c - L$$

$$S_{sa}^2 \quad .. \quad \sum_{j} \left(\sum_{s,\,i} \frac{(n_{ij}^{s})^2}{n_{i..}^{s}} - \sum_{s} \frac{(n_{.j}^{s})^2}{n_{..}^{s}}\right)\left(1 - \frac{1}{M_j}\right) \sigma_{j\alpha}^2/l - L$$

$$S_{so}^2 \quad .. \quad \sum_{j=1}^{n} \left(n_{.j} - \sum_{s} \frac{(n_{.j}^{s})^2}{n_{..}^{s}}\right)\left(1 - \frac{1}{M_j}\right) \sigma_{j\alpha}^2/L\,(m-1)$$

$S_{(sc)o}^2$ .. No addition

$$S_s^2 \quad .. \quad \sum_{j}\left[\sum_{s} \frac{(n_{.j}^{s})^2}{n_{..}^{s}} - \left(\frac{(n_{.j}^{s})^2}{n}\right) - \frac{1}{M_j}\left(\sum_{s} \frac{(n_{.j}^{s})^2}{n_{..}^{s}} - \frac{(n_{.j})^2}{n}\right]\right] \sigma_{j\alpha}^2/L - 1$$

$$S_p^2 \quad .. \quad \sum_{j}\left[\sum_{s}\left(\frac{(n_{.j}^{s})^2}{n_{.j}} - \frac{(n_{.j}^{s})^2}{n}\right) - \frac{1}{M_j}\left(n_{.j} - \frac{n_{.j}^{2}}{n}\right)\right] \sigma_{j\alpha}^2/m - 1$$

4.4. We now turn to the special cases,

*Case I.*—We shall give the expression for the gain in efficiency by using $m$ parties instead of one enumerating the same number of observations. Under the conditions of 4·2§ it is given by

$$\frac{\frac{m-1}{m^2}\sum_j \sigma_{j\alpha}{}^2\left[\sum_r \frac{(n_{..}^{s})^2}{n^2} - \frac{1}{M_j}\right] + \sum_{j=1}^{m}(\alpha_j - \bar{\alpha})^2/m}{-\frac{\sigma^2}{N} + \frac{\sigma^2}{n} + \frac{\bar{\sigma}_\delta{}^2}{n} + \frac{\sigma_\epsilon{}^2}{n} + \frac{1}{m^2}\sum_j\left(\frac{(n_{..}^{s})^2}{n^2} - \frac{1}{M_j}\right)\sigma_{j\alpha}{}^2 + \bar{\alpha}^2}$$

where

$$\bar{a} = \sum_{j=1}^{m} \alpha_j/m$$

and if estimated by

$$\frac{(m-1)\,[S_p{}^2 - S_{e0}{}^2]}{S_{e0}{}^2}$$

provides a upper limit for the efficiency since we are neglecting a positive term

$$-\frac{\sigma^2}{N} + \frac{1}{m^2}\left[\sum_{j=1}^{m}\left[\frac{\sum_s (n_{..}^{s})^2}{n^2} - \frac{1}{M_j}\right]\sigma_{j\alpha}{}^2 + \bar{\alpha}^2\right.$$

in the denominator. The same expression provides an upper limit for the gain in efficiency in the other cases II and III.

4.5. Lastly we consider the situation where $mL$ enumerators for the $m$ parties are randomly selected out of a population of $M$ enumerators available with mean $\alpha$ and variance $\sigma_\alpha{}^2$ and are also randomly split up into $m$ parties. Here $\alpha_j^{s}$ is one of the $M$ possible values $\beta_t$ $(t = 1, 2, \ldots, M)$. Here the expectations of the means are given by the following expressions:

$$\begin{aligned} E(y_{.j}^{s}) &= \mu_s + \alpha \\ E(y_{..}^{s}) &= \mu_s + \alpha \\ E(y_{.j.}) &= \frac{\sum_s \mu_s\, n_{.j}^{s}}{a} + \alpha \\ E(y_{...}) &= \frac{\sum_{s=1}^{L} \mu_s\, n_h^{s}}{n} + \alpha \end{aligned} \qquad (22)$$

and for the variances of the different means, we add

$$\sigma_\alpha{}^2\left(1 - \frac{1}{M}\right),\ \sigma_\alpha{}^2\left(\sum_j \frac{(n_{.j}^{s})^2}{(n^{s})^2} - \frac{1}{M}\right),\ \sigma_\alpha{}^2\left[\frac{\sum_s (n_{.j}^{s})^2}{n_{.j}{}^2} - \frac{1}{M}\right]$$

and

$$\sigma_\alpha^2 \left[ \frac{\sum_{s,j} (n_{.j}^s)^2}{n^2} - \frac{1}{M} \right]$$

to the successive equations in (16). The expected values of the mean squares can be obtained from that of the fixed enumerators case by changing $\alpha_j^s$ to $\alpha$ and adding a term to certain mean squares as follows:

$S_w^2$ .. No addition

$$S_{se}^2 \quad .. \quad \sigma_\alpha^2 \left[ n - \sum_{s,j} \frac{(n_{.j}^s)^2}{n_{..}^s} \right] \Big/ lc - L$$

$$S_{so}^2 \quad .. \quad \sigma_\alpha^2 \sum_{s,j} \left[ \sum_i \frac{(n_{ij}^s)^2}{n_{i.}^s} - \frac{(n_{.j}^s)^2}{n_{..}^s} \right] \Big/ l - L$$

$$S_{se}^2 \quad .. \quad \sigma_\alpha^2 \left[ n - \sum_{s,j} \frac{(n_{.j}^s)^2}{n_{..}^s} \right] \Big/ L(m-1)$$

$S_{(se)_0}^2$ .. No addition

$$S_s^2 \quad .. \quad \sigma_\alpha^2 \sum_{s,j} \left[ \frac{(n_{.j}^s)^2}{n_{..}^s} - \frac{(n_{.j}^s)^2}{n} \right] \Big/ L - 1$$

$$S_p^2 \quad .. \quad \sigma_\alpha^2 \sum_{s,j} \left[ \frac{(n_{.j}^s)^2}{n_{.j}} - \frac{(n_{.j}^s)^2}{n} \right] \Big/ m - 1$$

It may be seen that the mean squares can be obtained from the corresponding expressions in (21) by omitting terms containing $M_j$ and replacing $\sigma_{j\alpha}^2$ by $\sigma_\alpha^2$. $S_p^2$, of course, will be independent of $\mu$'s if

$$\frac{n_{.j}^s}{n_{.j}} = \frac{n_{..}^s}{n_{..}}$$

The efficiencies in special cases I, II and III are measured by

$$(m-1) \frac{[S_p^2 - S_{(se)_0}^2]}{S_{(se)_0}^2}$$

when we neglect

$$\sigma_\alpha^2 \left[ \sum_{s=1}^{L} \left( \frac{(n_{..}^s)^2}{mn^2} - \frac{1}{M} \right) \right] + \alpha^2 - \frac{\sigma^2}{N}$$

in the denominator.

It is interesting to compare the relative merits of the situations when the several parties can be identified with different schools of training and/or different methods of collecting data against the one

where the parties cannot be so identified and are formed *ad hoc* randomly out of an available pool of enumerators. The absolute average reduction in the variance of the mean obtained by employing $m$ parties instead of one is more in the case of the former situation as compared to the latter. The extra reduction is given by

$$\frac{m(L-1)}{mL-1}(L\sigma_b^2 - \sigma_\omega^2),$$

where $\sigma_b^2$ is the variance of the mean biases of the parties given by

$$\sum_j \left(\frac{\sum a_j^s}{L} - \frac{\sum_{s,j} a_j^s}{mL}\right)^2 \Big/ m-1$$

and

$$\sigma_\omega^2 = \sum_s \sum_j (a_j^s - a_j)^2/m(L-1).$$

The implication is that whenever replication is desirable in sample surveys, it can be used to better advantage if the different replications can be definitely identified with different agencies as for example in crop surveys of the I.C.A.R. where a part of the work is entrusted to the agricultural department and partly to the revenue department rather than divide the work *ad hoc* but independently among different groups out of the same agency collecting information by similar methods, as in the latter case $\sigma_b^2$ will be of the same order as $\sigma_\omega^2$.

## 5. Examples

We shall now give some examples of sample surveys conducted by the Indian Council of Agricultural Research to form an idea of the magnitude of the error components due to biases relative to the total error.

*Example* 1.—This is taken from the yield survey for estimating the average yield of wheat conducted in Sind in 1945–46.[2] The design corresponds to case I except that the sample sizes for the different enumerators are unequal and each selected unit is further sub-sampled. It may be interesting to recall the situations which led to the adoption of this design before we discuss the results. This arose because of the age-old practice in India of obtaining two independent estimates one from agricultural agency and the other from the revenue agency with a view to gain more confidence in the pooled result. In case the two estimates widely differed, further enquiries used to be made before striking a final mean. When, therefore, sample surveys on a probability basis for estimating the average yield were initiated on

a State-wide scale in India in 1943–44, this practice of organising the work in two independent samples was advisedly continued by the I.C.A.R. There were also additional advantages in adopting this practice. Administrators in India were sceptical of the merits of sample surveys on a probability basis in providing unbiased and reliable estimates of the average yield and some of them even opposed the introduction of these surveys as a substitute for the official method in vogue. The agreement between yield estimates obtained from two independent agencies thus served to create confidence in the administrators in the new method. There was also reluctance on the part of administrators to overburden their staff with more number of experiments which they thought would be required under the new procedure. The distribution of work between the two agencies was, therefore, necessary in order that the requisite number of experiments may be conducted without overburdening either of them. Thus in every stratum, usually a tehsil, the work was divided in two independent samples, one to be carried out by a locally posted member of the department of agriculture and the other by the local official of the department of revenue. We will first deal with the data relating to one stratum only, *viz.*, Tehsil Kamber, Larkana District. The design adopted for the survey was multistage sampling with village as the

TABLE I

*Yield Surveys on Wheat: Tehsil Kambar, District Larkana*

Estimates of Average Yield and Analysis of Variance

| | Revenue | Agriculture | Combined |
|---|---|---|---|
| Mean yield in chh/plot .. | 100·3 | 54·9 | 85·6 |
| Number of experiments .. | 25 | 12 | 37 |

Analysis of Variance in (Chh/plot)²

| Source | d.f. | Mean square |
|---|---|---|
| Between Enumerators .. | 1 | 16714·6 |
| Between villages within enumerators .. | 15 | 7377·8 |
| Within villages .. | 20 | 315·4 |

$\hat{\sigma}_{e\omega}^2 = 3306{\cdot}4$; $\hat{\sigma}_{\alpha}^2 = 468{\cdot}3$; $\hat{\sigma}_{\alpha}^2 + \hat{\sigma}_{e\omega}^2 + \hat{\sigma}_{of}^2 = 4090{\cdot}1$

primary unit, a field under wheat as the sub-unit with a plot of 1/40th of an acre as the ultimate unit of sampling. The sample assigned to the Revenue official consisted of 13 villages with two fields in each village and one plot in each field. The number of villages assigned to the agricultural official was four but the number of fields per village was three, giving a total of 12 plots. All except one experiment assigned to the Officers were carried out. Estimates of the average yield and of the different components of variation due to bias are shown in Table I. Because of the asymmetric design, $\sigma_\alpha^2$ and $\sigma_{e_0}^2$ had to be estimated by the following formulæ:

$$S_{e_0}^2 = S_{0f}^2 + \frac{\sigma_{e_0}^2}{n-m}\left\{f - \sum_{j=1}^{2}\frac{\Sigma f_{ji}^2}{f_j}\right\}$$

$$S_e^2 = S_{0f}^2 + \frac{\sigma_{e_0}^2}{m-1}\left\{\sum_{j=1}^{2}\frac{\Sigma f_{ji}^2}{f_j} - \frac{\Sigma\,\Sigma f_{ji}^2}{f}\right\}$$

$$+ \frac{\sigma_\alpha^2}{m-1}\left\{f - \frac{\Sigma f_j^2}{f}\right\},$$

where

$\sigma_{e0}^2 = \sigma^2 + \tilde{\sigma}_\delta^2 + \sigma_\epsilon^2$,

$S_{0f}^2$ = mean square within villages,

$S_{e0}^2$ = mean square between villages within enumerators,

$S_e^2$ = mean square between enumerators,

$\sigma^2$ = true variance between villages,

$f$ = number of fields in the tehsil,

$f_j$ = number of fields for the $j$-th enumerator,

$f_{ji}$ = number of fields in the $i$-th village for the $j$-th enumerator,

$n$ = number of villages in the tehsil, and

$m$ = number of enumerators, two in the present case.

The table shows that the two yield estimates differ very considerably. Part of the difference is due to the differential bias of the two enumerators and part due to other fluctuations including sampling. A test of significance of the differential bias is provided by the ratio of $S_e^2$ to $S_{e0}^2$. The design being non-orthogonal, however the two mean squares are not independent and the ratio of the two cannot be

regarded to follow the $F$ distribution. An appropriate test may be provided by a comparison of $S_e^2$ with

$$S_{e0}^2 \frac{k_1}{k_1'} + S_{0f}^2 \left(1 - \frac{k_1}{k_1'}\right)$$

where

$$k_1 = \left[f - \sum_{j=1}^{2} \frac{\Sigma f_{ji}^2}{f_j}\right] \Big/ (n - m)$$

$$k_1' = \left[\sum_{j=1}^{2} \frac{\Sigma f_{ji}^2}{f_j} - \frac{\Sigma\Sigma f_{ji}^2}{f}\right] \Big/ (m - 1)$$

The ratio can be regarded to be distributed as $F$ with the degrees of freedom given by

$$\frac{\left\{\left(1 - \frac{k_1}{k_1'}\right) S_{0f}^2 + \frac{k_1}{k_1'} S_{e0}^2\right\}^2}{\left(1 - \frac{k_1}{k_1'}\right) S_{0f}^4 + \left(\frac{k_1}{k_1'}\right)^2 S_{e0}^4}$$

On reference to $F$ tables, it will be found that the observed ratio is maller than the $F_{5\%}$ value showing that $\sigma_\alpha^2$ is not significant though ts estimate works out a little more than 10% of the total variation.

Now we consider the data relating to all the tehsils in the Larkana District. The enumerators in the different tehsils (strata) were different, one belonging to the Department of Revenue and the other to the Department of Agriculture. The results are given in Table II. The mean square between enumerators within strata has not been split into two parts as between the two parties (*i.e.*, agencies) and within the parties owing to the heavy calculations involved in the fitting of the constants and also because it was not considered worthwhile to undertake these computations seeing that the mean square between enumerators within tehsils was not statistically significant. The average magnitude of $\hat{\sigma}_\alpha^2$ over all strata has, therefore, been taken to represent the component due to the differential bias among the enumerators. It will be seen that $\sigma_\alpha^2$ is not significant though it works out a little less than 15% of the total variation affecting an observation.

Results of yield survey for one more district, *viz.*, Sukkur, where the work in each of the several strata was carried out by two independent agencies, are also given in Table II. The table shows that $\sigma_\alpha^2$ is not significant, the contribution accounted for by it being less than 2% of the total variation. The above results, which are typical of the yield surveys conducted by the Indian Council of Agricultural Research, indicate the possible absence of the differential bias among enumerators.

TABLE II

*Yield Surveys on Wheat in Larkana and Sukkur Districts*

Estimates of Average Yield and Analysis of Variance

| | LARKANA | | | SUKKUR | | |
|---|---|---|---|---|---|---|
| | Revenue | Agricul-ture | Combin-ed | Revenue | Agricul-ture | Combin-ed |
| Mean yield in ch/plot | 105·3 | 101·7 | 103·7 | 157·0 | 143·1 | 152·9 |
| Number of experiments. | 58 | 46 | 104 | 80 | 34 | 114 |
| | Analysis of variance in $(\text{ch/plot})^2$ | | | | | |
| Source | d.f. | M.S. | | d.f. | M.S. | |
| Between tehsils | 6 | 11719·8 | | 7 | 54530·7 | |
| Between enumerators within tehsils, | 7 | 9164·4 | | 8 | 14592·8 | |
| Between villages within enumerators, | 32 | 5622·2 | | 37 | 11275·8 | |
| Within villages | 58 | 2750·8 | | 61 | 4586·0 | |
| | | $\sigma_{\ell 0}^2$ = 1311·9<br>$\sigma_a^2$ = 493·8<br>$\sigma_a^2+\sigma_{e0}^2+\sigma_{0f}^2$ = 4556·5 | | | $\sigma_{e0}^2$ = 3301·3<br>$\sigma_a^2$ = 158·2<br>$\sigma_a^2+\sigma_{e0}^2+\sigma_{0f}^2$ = 8045·5 | |

*Example* 2.—The data for this example is derived from a pilot survey for comparing the relative efficiency of the different size plots in estimating the average yield of irrigated wheat in Moradabad District during 1945.[3] Two parties of enumerators, called *A* and *B* consisting of five enumerators each were put on the job. The two parties were formed *ad hoc* out of the statistical investigators of the Indian Council of Agricultural Research. One enumerator from each party worked in one tehsil of the district. In each tehsil, two independent samples of two villages each were selected and allotted one each to an enumerator. Subsampling of the villages was done as in example 1. The sizes of the different plots ranged from 12·5 to 471·6 sq. ft. The data relating to the plot size 471·6 sq. ft. (33′ △) forms the basis for our illustration. The different averages and the pooled analysis of variance are given in Table III.

TABLE III (*a*)

*Yield Surveys on Wheat in Moradabad District*

Estimates of Average Yield of Wheat ch/plot and Pooled Analysis of Variance

| | | Tehsil 1 Average | Tehsil 2 Average | Tehsil 3 Average | Tehsil 4 Average | Tehsil 5 Average | Tehsil 6 Average | District Average |
|---|---|---|---|---|---|---|---|---|
| Irrigated | *A* .. | 8 429·1 | 8 326·0 | .. .. | 8 220·5 | 8 452·0 | 8 586·2 | 40 402·6 |
| | *B* .. | 8 530·0 | 8 207·8 | .. .. | 6 367·2 | 8 304·0 | 8 348·7 | 38 350·7 |
| | Pooled .. | 16 479·6 | 16 266·9 | .. .. | 14 283·4 | 16 377·5 | 16 467·5 | 78 377·3 |
| Unirrigated | *A* .. | 8 215·0 | 8 104·9 | 16 224·9 | 8 310·9 | 8 241·2 | 8 300·4 | 56 231·7 |
| | *B* .. | 8 289·6 | 8 136·9 | 15 219·3 | 8 173·9 | 8 231·7 | 8 296·2 | 51 219·0 |
| | Pooled .. | 16 252·3 | 16 120·9 | 31 222·2 | 16 244·9 | 16 236·5 | 12 299·0 | 107 225·7 |

TABLE III (*b*)

Pooled Analysis of Variance in (ch/plot)$^2$

| Source of Variation | Irrigated | | Unirrigated | |
|---|---|---|---|---|
| | d.f. | M.S. | d.f. | M.S. |
| Between tehsils .. | 4 | 154038 | 5 | 42384 |
| Between enumerators within tehsils | 5 | 96490 | 6 | 24092 |
| Between villages within enumerators | 10 | 52299 | 15 | 74845 |
| Between fields within villages .. | 19 | 22832 | 27 | 16491 |
| Between plots within fields .. | 39 | 8482 | 53 | 5692 |

A glance at the table for irrigated wheat shows that there are rather large and varying differences between the averages of the *A* and *B* samples in the different tehsils. If $\alpha_A^s$, $\alpha_B^s$ are the average biases of the enumerators working in the *s*-th tehsil in the party *A* and *B* respectively, then the hypothesis

$$\sum_{s=1}^{t} (\alpha_A^s - \alpha_B^s)^2/L = 0$$

tests that $\alpha_A^s = \alpha_B^s$ for all *s*, *i.e.*, there is no differential average bias between the parties. This is tested by the ratio:

$$\frac{\text{mean square between enumerators within tehsils}}{\text{mean square between villages within enumerators}}$$

which is equal to 1·8. The value of this ratio shows the absence of the differential bias even though the magnitude of the average $\sigma_\alpha^{2s}$ is as high as 30% of the total variation of an observation.

Results of the survey on unirrigated wheat carried out by the same parties in the same district are given in the lower part of Table III. These results show again the possible absence of differential bias among the enumerators. The estimate of the average value of $\sigma_\alpha^{2s}$ is, in fact, zero.

*Example* 3.—This example relates to a socio-economic survey conducted by the students of the International Training Centre on Censuses and Statistics for South-East Asia held at I.C.A.R. during December 1949.[4] The survey was carried out in three villages: Badli, Shamapur and Auchandi, situated at a distance of 10 to 15 miles from Delhi. The houses in each village were serially numbered and grouped into blocks of three. A certain number of these blocks was selected at random and within each block alternate households, *i.e.*, families were enumerated. The sample for each village was divided into independent samples one each to be enumerated by a different party of students. Thus the work in the village Badli was divided among six parties of enumerators, that in Shamapur and Auchandi among four and two parties respectively. The questionnaire used for the survey was prepared by the students themselves and included a large number of items. The results given below, however, relate only to three characters, *viz.*, sex proportion, illiteracy proportion and the proportion of persons economically independent in a family. Table IV gives the estimated value for each of the three characters. The study of this table shows that there is more variability in the estimates given by different parties in the character, 'economic independence' than in the estimates of the other two characters, thereby suggesting that the contribution of the components due to differential bias of the parties is relatively more important in the case of economic independence. The estimated values of the components due to the differential bias of the parties, *i.e.*, $\sigma_\alpha^2$ and the remainder of the total variation, *i.e.*, $\sigma_\delta^2 + \sigma_\epsilon^2 + \sigma^2$ are given in Table V. The results of the test of significance show that $\sigma_\alpha^2$ is not significantly different from zero. The relative magnitude of $\sigma_\alpha^2$ as compared to the total variation is larger in the case of 'economic independence' than in the case of the other two characters. In one case it accounts for nearly 50% of the total variation.

TABLE IV

*Socio-Economic Survey in Delhi Villages*

Estimated Percentages of the Three Characters by Different Parties

| Character \ Party | Village Shamapur | | | | Village Auchandi | | Village Badli | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | I | II | I | II | III | IV | V | VI |
| Sex | 58·1 | 56·0 | 58·2 | 56·9 | 48·9 | 54·1 | 69·9 | 55·0 | 67·8 | 34·2 | 60·4 | 46·8 |
| Economic Independence | 61·5 | 27·1 | 39·1 | 44·8 | 26·1 | 47·6 | 67·1 | 45·8 | 76·7 | 49·6 | 46·3 | 31·1 |
| Illiteracy | 76·5 | 65·7 | 75·1 | 73·0 | 85·9 | 88·8 | 88·1 | 75·1 | 89·8 | 87·0 | 62·4 | 94·6 |

TABLE V

*Socio-Economic Survey in Delhi Villages*

Estimates of the Different Components of Variation

| Village | Character | $\sigma^2+\sigma_\delta^2+\sigma_\epsilon^2$ | $\sigma_a^2$ |
|---|---|---|---|
| Badli .. | Sex | 478·7 | 68·2 |
| Shamapur .. | | 390·9 | 0·0 |
| Auchandi .. | | 190·0 | 16·2 |
| Badli .. | Economic Independence | 722·1 | 103·8 |
| Shamapur .. | | 216·8 | 204·6 |
| Auchandi .. | | 487·0 | 0·0 |
| Badli .. | Illiteracy | 574·5 | 0·0 |
| Shamapur .. | | 552·5 | 38·6 |
| Auchandi .. | | 491·9 | 0·0 |

*Example* 4.—This example relates to the data collected in the course of a surprise check on acreage statistics maintained by the village patwaris.[5] Acreage under crops in the temporarily settled areas in India is compiled by the patwaris by noting the names of the crops field by field in the course of their normal administrative work. As all the fields are surveyed and mapped and the area of each field survey number) is, therefore, accurately known, the total area under

any crop is obtained by simply adding the area of the fields growing that crop.

This method, though sound in principle, is not always free from errors in practice. Indeed it is claimed that the patwaris do not exercise sufficient care in ascertaining the names of crops growing in the fields under their jurisdiction. A surprise check, was, therefore, organised in randomly selected villages of Lucknow District in Uttar Pradesh to examine the extent of the inaccuracy of the records maintained by the patwaris. The check was carried out by the staff of the statistical section of the Department of Agriculture, U.P. and the Indian Council of Agricultural Research.

Altogether 61 villages were selected at random for the purpose of this check. In each village 8 survey numbers were selected at random. In each survey number the statistical investigator was asked to record the name of the crop. In case more than one crop was grown, he was asked to give the names of the crops together with the proportion of the area under each. The patwari's records for the same survey numbers were available in the khasra book maintained by him. The check was carried out at harvest time after the patwaris had completed their inspection and made entries in the register. Tables VI and VII given below summarise the results. Table VI (*a*) classifies the survey numbers according as they were recorded to be under wheat or not and Table VI (*b*) gives the same information in respect of gram. The table shows that nonsampling errors due to misreportage are relatively small and suggest the conclusion that the

TABLE VI

*Classification of Khasra Numbers by Crops as Reported by the Two Agencies*

WHEAT

| Patwari Agency | Statistical Enumerator: Wheat | Statistical Enumerator: No Wheat | Statistical Enumerator: Total |
|---|---|---|---|
| Wheat .. | 121 | 4 | 125 |
| No Wheat .. | 5 | 358 | 363 |
| Total .. | 126 | 362 | 488 |

GRAM

Statistical Enumerator

| Patwari Agency | Gram | No Gram | Total |
|---|---|---|---|
| Gram .. | 111 | 3 | 114 |
| No Gram .. | 10 | 364 | 374 |
| Total .. | 121 | 367 | 488 |

differential biases among the enumerators are almost absent. We shall treat the example by the methods developed in this paper and estimate the differential bias between the two agencies.

Unlike previous examples, this example comes under case III. A village corresponds to a stratum here and $L = 61$, $n_{ij}^{s} = 0$ or 1, $n_{i.}^{s} = p = 2$, $l_s = 8$, $m = 2$, $l = 488$ and $n = 976$. Substituting the values in the expectations of different mean squares, we obtain

$$S_p^2 = 488 \sum_{j=1}^{2} (y_{.j} - y_{..})^2 \quad \text{estimates} \quad 488\,\sigma_b^2 + \sigma_\delta^2$$

$$S_{e0}^2 = \frac{\sum_{s=1}^{61} \sum_{j=1}^{2} \sum_{i=1}^{8} (y_{ij}^{s} - y_{.j}^{s})^2}{7 \times 2 \times 61} \quad \text{estimates} \quad \sigma^2 + \sigma_\delta^2$$

$$S_{s0}^2 = \frac{2 \sum_{s} \sum_{i} (y_{i.}^{s} - y_{..}^{s})^2}{7 \times 61} \quad \text{estimates} \quad 2\sigma^2 + \sigma_\delta^2,$$

where $y_{ij}^{s} = 0$ or 1 according as the field is under wheat (gram) or not. The values of $\sigma^2$, $\sigma_b^2$ and $\sigma_\delta^2$ calculated from the above are given below:

| | $\sigma^2$ | $\sigma_b^2$ | $\sigma_\delta^2$ |
|---|---|---|---|
| Wheat .. | ·1604 | 0·0000 | 0·0089 |
| Gram .. | ·1490 | 0·0001 | 0·0000 |

These results show that the intrinsic variation between the sampling units accounts for almost the whole of the variation and the component

of interaction between enumerators and the observations as also that due to differential bias between the two parties are negligible. The results confirm the observations made earlier, from an inspection of the two tables.

*Example* 5.—We will now give an example where we can estimate not only the component due to differential bias of the enumerators $\sigma_\alpha^2$ but also the contribution due to the intrinsic variation $\sigma^2$ of the units and that arising from observations round the average bias of the enumerators $\sigma_\delta^2 + \sigma_\epsilon^2$. Replication alone in the form of two or more independent samples cannot give separate estimates of $\sigma^2$ and $\sigma_\delta^2 + \sigma_\epsilon^2$ unless more than one measurement is made on a unit by different enumerators. Accordingly in a survey[6] conducted by the students of the Indian Council of Agricultural Research in 8 villages of the Delhi Development Scheme, provision was made for repeated

TABLE VII

| Source | D.F. | M.S. | Estimate of |
|---|---|---|---|
| Enumerators (eliminating units) | $m-1$ | $\frac{\Sigma b_j Q_j}{m-1}$ | $\sigma_\delta^2+\sigma_\epsilon^2+\frac{n-l}{m-1}\sigma_\alpha^2$ |
| Units (ignoring enumerator) | $l-1$ | $\frac{\sum_{i=1}^{l}\frac{T_i^2}{n_i.}-\frac{T^2}{n}}{l-1}$ | $\sigma_\delta^2+\sigma_\epsilon^2+\frac{n-\frac{\Sigma n_{.j}^2}{n}}{l-1}\sigma^2+\frac{l-\frac{\Sigma n_{i.}^2}{n_{.j}^2}}{l-1}\sigma_\alpha^2$ |
| Error | $n-m-l+1$ | $\frac{\Sigma\Sigma y^2-\sum_{i=1}^{l}\frac{T_i^2}{n_i}-\Sigma b_j Q_j}{n-m-l+1}$ | $\sigma_\delta^2+\sigma_\epsilon^2$ |

where

$T_i$ denotes the total of the observations made on the $i$-th unit.

$T$ denotes the total of all the observations and $b_j$ are given by

$$\sum_{j=1}^{m} c_{jj'} b_j = Q_j$$

in which

$$c_{jj} = n_{.j} - \sum_{i=1}^{l} \frac{n_{ij}^2}{n_{i.}}$$

$$c_{jj'} = -\sum_{i=1}^{l} \frac{n_{ij} n_{ij'}}{n_{i.}}$$

and

$$Q_j = T_{.j} - \sum_{i=1}^{l} \frac{T_i n_{ij}}{n_{i.}}.$$

$T_{.j}$ being the total of all the observations made by the $j$-th enumerator.

measurements on some units. Four grids of 4 khasra numbers were selected in each of the 8 villages. The plan followed was that each enumerator visits all the khasra numbers in the 4 grids subject to the restriction that one of the grids is just once reported on by him and the other three are reported one each in common with one and only one of the other enumerators. All the fields within a khasra number were measured, the character measured being $\frac{\text{breadth}}{\text{length}} \times 100$. A field was defined as a contiguous piece of land growing the same crop.

The method of analysis of variance was that for the incomplete block design. The enumerator corresponds to the block and the intrinsic value of the character of the field as a variety. The analysis of variance in Table VII provides the estimates of $\sigma^2$, $\sigma_\alpha{}^2$ and $\sigma_\delta{}^2 + \sigma_\epsilon{}^2$ besides giving the test of significance of $\sigma_\alpha{}^2$ being different from zero.

TABLE VIII

*Area Survey in Delhi Villages*

Estimates of the Components $\sigma^2$, $\sigma_\alpha{}^2$ and $\sigma_\delta{}^2 + \sigma_\epsilon{}^2$

| Name of village enumerated | $\sigma_\delta{}^2+\sigma_\epsilon{}^2$ | $\sigma_\alpha{}^2$ | $\sigma^2$ |
|---|---|---|---|
| Bawana .. | 33·2 | 0·0 | 614·0 |
| Sultanpur .. | 102·2 | 21·6 | 375·4 |
| Barwala .. | 48·4 | 4·7 | 332·8 |
| Prahladpur .. | 29·4 | 0·0 | 418·8 |

## 6. DISCUSSION OF THE RESULTS

We have seen that replicated samples provide us with a method of estimating non-sampling errors. We have also given examples from different fields to obtain an idea of the magnitude of these errors relative to the total error. Although in the examples we have given, non-sampling errors have been found to be relatively unimpotrant, there may be cases particularly in socio-economic surveys where they may be sufficiently large to vitiate the estimate and thus it may be desirable to know their extent. A question naturally arises whether replicated samples should be incorporated as a regular feature of sample surveys. Mahalanobis adopts this design as a regular feature of the sample surveys[7, 8] and U.N. Sub-Commission on Sampling[9] has emphasised its importance in surveys. This point has also previ-

ously been examined by several authors.[10, 11, 12] We will examine the same in the light of the results of this paper.

We have shown that $S_p^2/S_{e0}^2$ provides an over-all test for finding whether the average bias differs from party to party or not. For any single stratum, it takes the form $S_e^2/S_{e0}^2$ and tests the absence of differential bias among the enumerators working within a stratum. This latter test does not, however, have a good discriminating power unless the size of the sample allotted to the stratum is sufficiently large. Normally the size of the sample in any individual stratum cannot be large as the survey is intended mainly to estimate the average for the whole of the population rather than for individual strata. In consequence the test fails most of the time to reveal the significant existence of non-sampling errors even when the differences between the several estimates are large as, for instance, in Example 1 of this paper.

One method of improving the discriminating power of the test is to use linked samples. This helps to reduce the standard error of the differences between the samples. On the other hand the standard error of the pooled estimate is increased by $\frac{(m-1)\rho\sigma^2}{n}$ where $\rho$ is the inter-class correlation between the members in the linked samples with a corresponding loss in efficiency due to this procedure in relation to the one where the replicated samples are not so linked. Even when the results are pooled over all the strata put together, this limitation on the size of the sample in each stratum is not enitrely got over. The test of significance based on $S_p^2/S_{e0}^2$ will doubtless be more conclusive than the corresponding test for each stratum but the significance will not enable us to detect the reliability or otherwise of the field work of individual enumerators since it is just a test of the differences between average biases of the $m$ parties, each averaged over all strata. On the other hand, this test can show complete agreement even when there may be individual differential biases among enumerators. The test of significance based on $\sum_{s=1}^{L} S_{se}^2 \Big/ \sum_{s=1}^{L} S_{(se)0}^2$ does not suffer from this defect but it, like the test $S_p^2/S_{e0}^2$, can point out disagreement and for the location of the disagreement we have to go back to the results for the individual strata which, as stated before, will not reveal the differential bias most of the time. In other words, replicated samples can point out disagreement without telling us where to locate it. Mahalanobis suggests an alternative procedure to the over-all tests for locating disagreement. He calculates $F$ for each stratum and tabulates the frequency distribution of the $L$ values so obtained and

examines whether the distribution corresponds to the known distribution of $F$. In case more than 5% of the $F$ are significant, the disagreement is considered entirely due to the work in these strata and these strata alone are examined for finding possible causes of disagreements. We give below a table adapted from Mahalanobis showing the distribution of $F$ in comparisons of half samples $A$ and $B$ in 379 strata into which the population was divided. The results relate to the jute area survey in Bengal in 1941.

*Comparison of half-samples (A) and (B): Students' 't' for strata*

| Range of probability of $t$-values | Number of cases | | Difference | |
|---|---|---|---|---|
| | Observed | Expected | | |
| (1) | (2) | (3) | (4) | (5) |
| *Less* than 0·05 .. | 109 | 18·95 | +90·05 | 427·92 |
| 0·05–0·10 .. | 220 | 18·95 | + 1·05 | 0·06 |
| 0·10–0·90 .. | 235 | 303·20 | −68·20 | 15·34 |
| 0·90–0·95 .. | 12 | 18·95 | − 6·95 | 2·55 |
| 0·95–1·00 .. | 3 | 18·95 | −15·95 | 13·42 |
| Total .. | 379 | 379·00 | .. | 459·29 |

As the table shows $F$ is significant in 109 out of 379 strata. Mahalanobis, therefore, undertook a scrutiny of the field record in those 109 strata and found that in as many as 84 of these, the discrepancies could be ascribed to what he calls real physical differences between the two half-samples, $A$ and $B$, within each stratum. Omitting these 84 strata, he then tabulates the distribution of the remaining 295 values of $F$. He compares it with the expected distribution and finds that the two are in satisfactory agreement as only 25 out of 295 $F$ values are now significant. From this he concluded that the object of using the replicated (half) sampling method was entirely successful.

To us this claim for the usefulness of replicated samples does not appear to be entirely justified. For once the discrepant work is suspected, one should scrutinise the work in all the strata and not confine the scrutiny to those strata having significant $F$. For, in strata where $F$ is non-significant, we can also expect discrepant work,

as the non-significance can be due to the opposite effects of the discrepancy in work and the real physical differences between the samples $A$ and $B$. Again, when the sample size is small, as it will usually be in each stratum, the method of Mahalanobis may lead us to looking for trouble where it does not exist and *vice versa* since it is quite likely that real large differences may be declared non-significant and *vice versa*. Further in large sample surveys, particularly when the survey is over, it is not possible to go back to the units for the scrutiny required under replicated samples. This also brings out another difficulty in a lay out under replicated samples as the resulting sample may not represent the same population in all strata. In our opinion the whole procedure of accepting the verdict of agreement where $F$ is non-significant and explaining away the disagreement in terms of the physical differences where $F$ is significant is logically untenable.

Quite apart from the limitation of the size of the samples within each stratum which renders replicated samples an ineffective tool for detecting discrepancies in field work, there is another factor which needs to be conisdered before recommending replicated samples as a regular feature of sample surveys, *viz.*, the cost of the survey. The use of the replicated samples requires that each enumerator covers the entire stratum thereby adding to the travel component of the cost of the survey. Assuming that $m'$ enumerators, as against $m$ in replicated samples, are required in case replicated sample is not employed and the sample is divided into $m'$ compact groups, one such group to be enumerated by one enumerator, the cost of the survey on account of travelling will be only $\frac{1}{\sqrt{m}}$ of the corresponding cost for the replicated samples. This is on the assumption that the travelling cost per enumerator within an area of magnitude $A$ is proportional to $\sqrt{An}$, where $n$ is the number of units to be enumerated by each enumerator randomly distributed over the whole area. If travel cost forms a considerable part of the cost of a survey, as for instance in the case of surveys conducted by Mahalanobis with the help of an *ad hoc* staff, the consequent loss in efficiency can be appreciably large as shown by Mokashi.[13]

In view of the above, replicated samples cannot be recommended as a regular feature of sample surveys. However, the use of replicated samples may be considered when non-sampling errors are likely to be large but even here we would recommend its use at the pilot stage for improving the questionnaire and the method of training with a view

to reduce non-sampling errors rather than as an integral part of the large-scale surveys. If non-sampling errors cannot be so controlled by improving the questionnaire and training to the level of accuracy with which information is desired to be sought, one would hesitate to take up a sample survey on a probability basis.

Although replicated samples cannot by itself be an effective tool for the control of the field work, the need of controlling it in other ways is indeed urgent. In our view this need can be best met by the age old practice of providing adequate and effective supervision over the field work. It is, therefore, important to examine in what respects supervision differs from the replicated samples and how it provides a better control over the reliability of field work. Supervision differs from the usual form of replicated samples in many respects, *viz.*,

(i) It will be carried out by the superior staff, better paid, qualified and experienced as compared to the enumerators at the primary level employed in replicated samples.

(ii) It will be carried out on a part of the work performed at the primary level whereas replicated sample requires at least two independent samples.

(iii) Supervision is not confined only to enumerating for the second time units once observed at the primary level. It has a wider objective in view, namely that of correcting and improving the field work on the spot, whereas replicated sample will usually suggest the improvement when the survey is over.

(iv) A supervisor need not be present throughout the operations connected with the enumeration of a selected unit, whereas an enumerator under a replicated sampling scheme has got to enumerate completely every unit assigned to him.

(v) Units selected for supervision may or may not be selected by the principle of random sampling, whereas in replicated samples they will necessarily be so selected. When it is possible to arrange supervision on a probability basis and the work done by the supervisors is considered a sub-sample of the work done at the primary level, supervision may be considered a very special form of replicated samples subject to the differences mentioned above. This way supervision provides an external evidence against which to check the work at the primary level.

(vi) Replicated samples will not reveal minor defects in an investigator and will certainly not reveal faults which are common to all the investigators, whereas this is possible with supervisory checks.

(vii) Replicated samples alone can estimate non-sampling errors whereas supervision will not unless conducted as visualised in (v).

(viii) When supervision is carried out as in (v), the results of the supervision can be utilised to improve the estimates obtained from the work done at the primary level.

It follows that supervision can provide a better control over field work in a variety of ways which is not possible in case of replicated samples. Replicated sample is no alternative to supersivory check, though the latter can be. Replicated sample has a place when either the object of the survey is to compare between different methods or different classes of investigators or at the pilot stage of a large-scale survey but it is effective only when accompanied by adequate supervision.

## 7. Summary

This paper deals with the measurement of non-sampling errors by the method of replicated samples. Formulæ have been developed to give the estimates of the various components of non-sampling errors. Simple special cases of the replicated samples have been discussed with illustrations from the surveys carried out under the auspices of the Indian Council of Agricultural Research. Lastly, the limitations of the method of replicated samples as a means of controlling the quality of field work by making it an integral feature of the sample surveys have been pointed out and the relative merits of the method of replicated samples and supervision examined.

## 8. References

1. Sukhatme, P. V., "The Problem of Plot Size in Large-scale Yield Surveys," *Jour. Amer. Stat. Assoc.*, 1947 *a*, **47**.

2. Indian Council of Agricultural Research, New Delhi, "Report on crop-cutting experiments for estimating the average yield of wheat, Sind," 1945–46.

3. Indian Council of Agricultural Research, New Delhi, "Report on crop-cutting experiments for estimating the average yield of wheat, U.P.," 1944–45.

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.