POST CLUSTER SAMPLING[1]

By Sakti P. Ghosh[2]

University of California, Berkeley

1. Introduction. The main difficulty often faced in cluster sampling is the lack of information relating to the composition of the clusters. In such situations the clusters can be built up on the basis of an initial random sample, then the final sampling can be done with these clusters as sampling units. Thus the name "Post Cluster", which essentially means that the clusters are formed afterwards and are not known beforehand, has been coined by T. Dalenius, who first introduced the idea in his book *Sampling in Sweden*, pp. 156–158.

The notion of forming groups of units (which may either be regarded as strata or clusters) after drawing an initial sample is due to Friedman and Wilcox. They queried Neyman at a Conference on Sampling Human Populations held in Washington, D. C., in April 1937 whether there was a solution to the problem of the optimum size of the initial sample and the smaller stratified sample selected from the initial sample. Neyman solved this problem subsequently and gave the results in a paper in 1938. In the Friedman-Wilcox method a certain number of units (indicated by the relevant optimum allocation theory) are sampled from every cluster (or strata); but in the method proposed in the paper a subset of entire clusters is selected. Later the Friedman-Wilcox method came to be known as phase sampling. David following Neyman (1938), among other things, gave a slightly more general treatment of the problem and also used the method of characteristic random variables. The common feature in the two methods is the regrouping of units after initial sampling according to some rule.

This technique of sampling differs from phase sampling in that it uses a hierarchy of sampling units, and differs from ordinary subsampling in that sampling units at the second stage are larger than the sampling units at the first stage.

The purpose of this paper is to develop a stochastic model for analysis of sampling problems that may arise in cluster sampling when the composition of the clusters is not at hand.

2. A stochastic model for the selection of the initial random sample. The type of sampling outlined above, will obviously depend on the "rule" which will be adopted in forming the clusters out of the sampled elements. We shall first consider the simplest situation where the clusters already exist in the population but the elements cannot be identified to the proper clusters until some auxiliary character is observed. The model behind the sampling procedure can

Received March 14, 1962; revised November 2, 1962.

[1] This paper was prepared with the partial support of the Office of Ordnance Research, U.S.A., Grant (DA-ARO(D)-31-124-G183).

[2] Now at Thomas J. Watson Research Center of IBM.

be looked at from the point of view of an urn-scheme as follows: We have a large urn containing N balls. These balls are of M different colors, but the colors cannot be identified unless the balls are drawn. Now at the first stage of the sampling procedure we draw n balls out of this urn and place them into M small urns according to the color of the balls (some urns may be empty). At the second stage we shall select m small urns from these M small urns and then measure some properties of the balls, say diameter, etc.

The statistical model which will be the basis of the analysis may be described as follows. Suppose we have a finite population (π) consisting of N distinct elements and the values, which the character under analysis can take, are given by $X_1', X_2', \cdots, X_N'$. These N elements may in principle be grouped into $M(> 1)$ clusters, which may be denoted by C_i's, $i = 1, 2, \cdots, M$ and C_i contains N_i elements, i.e., $\sum_{i=1}^{M} N_i = N$. The elements of C_i are given by $C_i = (X_{i1}, X_{i2}, \cdots, X_{iN_i})$ where $(i1, i2, \cdots, iN_i) \varepsilon (1, 2, \cdots, N)$. We are interested in estimating the mean per element, i.e.,

$$\bar{X} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}/N.$$

Suppose we first draw a random sample of n elements without replacement. Let us introduce a set of random variables $\epsilon_{ij}, j = 1, 2, \cdots, N_i, i = 1, 2, \cdots, M$, indicating the performance of the random sample, i.e.,

$\epsilon_{ij} = 1$ if X_{ij} is selected in the random sample and this event can happen with a probability $= n/N$,
$= 0$ otherwise.

Hence our random sample selected is given by

$$R_1 = (\epsilon_{11}X_{11}, \epsilon_{12}X_{12}, \cdots, \epsilon_{MN_M}X_{MN_M}).$$

The clusters formed out of R_1 will be denoted by d_i's and are given by

$$d_i = (\epsilon_{i1}X_{i1}, \epsilon_{i2}X_{i2}, \cdots, \epsilon_{iN_i}X_{iN_i}) \qquad i = 1, 2, \cdots, M.$$

3. Selection of post clusters with equal probabilities. In this section we will primarily be concerned with the selection of m clusters with equal probabilities without replacement. Here also we introduce another set of random variables $\phi_i, i = 1, 2, \cdots, M$ indicating the performance of selection of clusters, i.e.,

$\phi_i = 1$ if d_i is selected in the sample and this can happen with probability m/M,
$= 0$ otherwise.

The ϕ_i's are dependent among themselves but because of independence between sampling of elements and sampling of clusters (empty clusters are also to be selected—discussed in more detail in the footnote[3]). However ϕ_i's and the

[3] In actual realization any of the d_i's may be identically zero. In such cases also, we shall assume while drawing samples from the d_i's that the particular d_i exists hypothetically. This presents no difficulty in building up the estimate if $d_i \equiv 0$ appears in the sample and at the same time retains stochastic independence of the ϕ_i and ϵ_{ij}'s.

ϵ_{ij}'s are stochastically independent. Hence the sample finally selected out of π can be denoted by $S = (\phi_1 d_1, \phi_2 d_2, \cdots, \phi_M d_M)$.

Unbiased estimate of $\bar{X}$. Without any confusion we can denote the total of the values of the X_{ij}'s in d_i by the same symbol, i.e., $d_i = \sum_{j=1}^{N_i} \epsilon_{ij} X_{ij}$, hence $\phi_i d_i = \phi_i \sum_{j=1}^{N_i} \epsilon_{ij} X_{ij}$. Thus the sample total is given by

$$x_1 = \sum_{i=1}^{M} \phi_i d_i = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \phi_i \epsilon_{ij} X_{ij}.$$

LEMMA 1. *$\bar{x} = Mx_1/mn$ is an unbiased estimate of $\bar{X}$.*

PROOF. On account of the independence of the ϕ_i's and the ϵ_{ij}'s we have

$$E(x_1) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij} E(\phi_i) E(\epsilon_{ij}) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij} \frac{m}{M} \cdot \frac{n}{N} = \frac{mn}{M} \bar{X}.$$

Thus the lemma follows immediately.

REMARK 1. We have assumed that the ϕ_i's are independent of the ϵ_{ij}'s but when the zero d_i are omitted, while selecting a sample from the d_i's, some dependency is introduced, but still an unbiased estimate can be developed for the situation as follows: Suppose ν (random number) of d_i's are not identically zero. Then we define

$\phi_i = 1$ if d_i is selected in the sample with probability $= m/\nu$ $(\nu > m)$,
$= 0$ otherwise.

Thus $\phi_i = 0$ also for those i for which $d_i \equiv 0$. Now our estimate can be defined as

$$\bar{x}_2 = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \phi_i \epsilon_{ij} X_{ij} \nu / mn.$$

$\bar{x}_2$ will be unbiased because $E(\bar{x}_2) = E_{R_1} E(\bar{x}_2/R_1)$ where $E(\bar{x}_2/R_1)$ is the conditional expectation of $\bar{x}_2$ given R_1 and E_{R_1} is the expectation over R_1. On simplification $E(\bar{x}_2) = \bar{X}$. In this case, however, the variance will depend on ν and this may present some difficulty.

LEMMA 2. *If X and Y are stochastically independent, then*

$$V(X \cdot Y) = E(Y^2) V(X) + E^2(X) V(Y) = E(X^2) V(Y) + E^2(Y) V(X).$$

This lemma is due to Quenouille (1958, p. 37) and subsequently given in a slightly different form by Goodman (1960).

LEMMA 3. *$E(\epsilon_{ij}) = n/N$; $V(\epsilon_{ij}) = n(N-n)/N^2$; $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{il}) = -n(N-n)/N^2(N-1)$ for $j \neq l$; $E(\phi_i) = m/M$; $V(\phi_i) = m(M-m)/M^2$; $\mathrm{Cov}(\phi_i, \phi_k) = -m(M-m)/M^2(M-1)$ for $i \neq k$.*

LEMMA 4. *If $\{X_1, X_2\}$ are independent of $\{Y_1, Y_2\}$ (but X_1 is dependent on X_2 and Y_1 is dependent on Y_2) then*

$$\mathrm{Cov}(X_1 \cdot Y_1, X_2 \cdot Y_2) = \mathrm{Cov}(X_1 X_2) E(Y_1 \cdot Y_2) + \mathrm{Cov}(Y_1, Y_2) E(X_1) E(X_2).$$

Lemmas 3 and 4 are obtained easily from definitions.

COROLLARY. $\mathrm{Cov}(X \cdot Y_1, X \cdot Y_2) = E(Y_1 \cdot Y_2) V(X) + E^2(X) \mathrm{Cov}(Y_1, Y_2)$.

Using Lemmas 2, 3, and 4 the variance of $\bar{x}$ can be calculated directly and

after simplification is given by

$$V(\bar{x}) = \frac{M(N-n)}{mn(N-1)}\sigma^2 + \frac{M^2(M-m)(n-1)}{Nmn(N-1)(M-1)}V_c + \frac{(M-m)(N-n)}{mn(N-1)}\bar{X}^2 \tag{1}$$

where

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}^2 - \bar{X}^2 = \text{Variance of the elements.}$$

$$V_c = \frac{1}{M}\sum_{i=1}^{M}\left(X_i - \frac{X}{M}\right)^2 = \text{Variance of the cluster totals.}$$

The details of the proof are given in the appendix.

It is of interest to investigate whether the estimate $\bar{x}$ based on post cluster sampling (PCS) is better than random sampling (RS) of elements without replacement. The comparison becomes a little difficult because of the fact that the number of elements in PCS is a random variable. So the only way the comparison can be made is to consider the expected number of elements. The expected number of elements in PCS is mn/M. The variance $V(S_R)$ of the mean of an RS without replacement of mn/M elements is given by

$$V(S_R) = (M/mn)[(N - (mn/M)/(N-1)]\sigma^2.$$

It is easy to see that $V(\bar{x})$ can be expressed as

$$V(\bar{x}) = V(S_R) + \frac{(M-m)}{m(N-1)}\left\{\frac{M^2(n-1)}{nN(M-1)}V_c + \frac{1}{N}\left(\frac{X^2}{n} - \sum_i\sum_j X_{ij}^2\right)\right\}. \tag{2}$$

The term in the second bracket of (2) enables us to state the situations when the unbiased estimate of $\bar{X}$ based on PCS will be better or worse than RS. Thus $V(\bar{x}) <, =, > V(S_R)$ accordingly as

$$[M^2(n-1)/n(M-1)]V_c <, =, > N\sigma^2 - X^2[(1/n) - (1/N)]. \tag{2'}$$

The sign of equality holds in the trivial situation when $m = M$.

In a random population, i.e. when the clusters have same variance and hence $V_c = 0$, we have $V(x) <, =, > V(S_R)$ accordingly as $\bar{X}(N/n - 1) <, =, > \sigma^2$.

In most practical situations $\sigma^2 > \bar{X}(N/n - 1)$ and hence for random populations random sampling would be better than PCS.

Ratio estimate of $\bar{X}$. The population mean $\bar{X}$, in case of PCS can be looked upon as the ratio $R = X/N$ of the expectation of two random variables, viz, the variable under analysis and the number of observations. Hence, a ratio estimate of $\bar{X}$ based on PCS can be stated in the form,

$$\hat{x} = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\phi_i\epsilon_{ij}X_{ij} \Big/ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\phi_i\epsilon_{ij}.$$

It is obvious that $\hat{x}$ will be a biased estimate, as it is in the case of ordinary ratio estimate. (An unbiased ratio estimate can also be obtained by modifying

the design a little as shown by Lahiri (1951) but we shall not discuss his design here.) An approximate expression for the bias can be developed by using the classical technique of expanding the denominator in Taylor series and then taking expectation. $\hat{x}$ is unbiased when we neglect terms of $o(1/N)$ and the first approximation to the magnitude of bias, i.e., when we neglect terms of $o(1/N^2)$ is given by

$$\text{E}(\hat{x}) - R = R[(M-m)(n-1)N/mn(M-1)(N-1)] \{C(N_iN_i) - C(N_iX_i)\}, \tag{3}$$

where $C(N_iN_i) = M^2V(N_i)/N^2$ = Square of the coefficient of variation of the N_i's, $C(N_iX_i) = \text{Cov}(N_i, X_i)/(XN/M^2)$ = Coefficient of covariation between X_i and N_i.

The details of the proof are given in the appendix.

For large N we can approximate $N/(N-1)$ by 1; thus we have

$$\text{Bias} = \frac{R}{m}\left(1 - \frac{1}{n}\right)\left(1 - \frac{m-1}{M-1}\right)[C(N_i N_i) - C(N_i X_i)]. \tag{4}$$

As m increase the bias decreases rapidly through both factors $1/m$ and $[1 - (m-1)/(M-1)]$ showing that this ratio estimate is a consistent estimate as m increases but not as n increases. This ratio estimate has the usual property of an ordinary ratio estimate that the bias vanishes when the regression of X_i on N_i is linear, i.e., $C(N_iN_i) = C(N_iX_i)$.

It would be of interest to look into the variance of the ratio estimate to have an idea of the precision of the estimate. The variance of the ratio estimate will be an approximate one, neglecting terms of $o(1/N^2)$. Since the ratio estimate is unbiased neglecting terms of $o(1/N)$, this approximate variance will be equal to the mean square error neglecting terms of $o(1/N^2)$

$$\begin{aligned} V(\hat{x}) &= E[\hat{x} - E(\hat{x})]^2 \\ &\approx R^2E\left[\frac{\sum_i\sum_j(\phi_i\epsilon_{ij} - mn/MN)X_{ij}}{(mn/MN)X} - \frac{\sum_i\sum_j(\phi_i\epsilon_{ij} - mn/MN)}{(mn/MN)N}\right]^2 \\ &= R^2\left[\frac{V(\sum\sum\phi_i\epsilon_{ij}X_{ij})}{(mn/MN)^2X^2} + \frac{V(\sum\sum\phi_i\epsilon_{ij})}{(mn/MN)^2N^2} \right. \\ &\qquad \left. - \frac{2\,\text{Cov}(\sum\sum\phi_i\epsilon_{ij}X_{ij}, \sum\sum\phi_i\epsilon_{ij})}{(mn/MN)^2NX}\right]. \end{aligned}$$

All these terms in the bracket have already been calculated, hence substituting, we have

$$V(\hat{x}) = R^2\left[\frac{1}{m}\left(1 - \frac{1}{n}\right)\left(1 - \frac{m-1}{M-1}\right)\left(\frac{N}{N-1}\right)(C(X_iX_i) - 2C(N_iX_i) + C(N_iN_i) + \frac{M(N-n)}{(N-1)mn}C(X_{ij}X_{ij})\right], \tag{5}$$

where $C(X_{ij}X_{ij}) = \sigma^2/\bar{X}^2$ and $C(X_iX_i) = V_c/(X/M)^2$.

Assuming $(N-1)/N$ approximately to be unity, we have

$$V(\hat{\bar{x}}) \approx R^2 \frac{1}{m}\left(1-\frac{1}{n}\right)\left(1-\frac{m-1}{M-1}\right)\left[C(X_i X_i) - 2C(N_i X_i) + C(N_i N_i) + \frac{(1-n/N)(M-1)}{(1-m/M)(n-1)} C(X_{ij} X_{ij})\right]. \tag{6}$$

Let us consider the reduction in variance, due to the use of ratio estimate, over the unbiased estimate. By simplifying (1) a little and then subtracting (6) from it, it follows that

$$V(\bar{x}) - V(\hat{\bar{x}}) = R^2 \frac{1}{m}\left(1-\frac{1}{n}\right)\left(1-\frac{m-1}{M-1}\right) \cdot \left[\frac{(M-1)}{(n-1)}\left(1-\frac{n}{N}\right) + 2C(N_i X_i) - C(N_i N_i)\right]. \tag{7}$$

Equation (7) also gives us a chance to find the situations when $V(\bar{x}) \geqq V(\hat{\bar{x}})$, namely when the quantity in the third bracket of (7) is nonnegative, i.e., $(M-1)(N-n)/(n-1)N + 2C(N_iX_i) - C(N_iN_i) \geqq 0$.

On simplification, the condition becomes

$$\beta(X_i/N_i) \geqq \frac{X}{2N}\left[1 - \frac{M-1}{n-1}\left(1-\frac{n}{N}\right)\frac{1}{C(N_i N_i)}\right], \tag{8}$$

where $\beta(X_i/N_i)$ is the regression coefficient of X_i on N_i.

For ordinary cluster sampling, the situation when the ratio estimate is better than the unbiased estimate is $\beta(X_i/N_i) \geqq X/2N$, hence from (8) it is obvious that for the PCS the ratio estimate may be better than the unbiased estimate even when the ratio estimate is not better than the unbiased estimate for ordinary cluster sampling.

Special cases.

(i) *The regression between X_i and N_i is linear.* In such cases $\hat{x}$ will be unbiased and then

$$V(\hat{\bar{x}}) = V(\bar{x}) - \frac{R^2}{m}\left(1-\frac{1}{n}\right)\left(1-\frac{m-1}{M-1}\right)\left[\frac{M-1}{N-1}\left(1-\frac{n}{N}\right) + C(N_i X_i)\right].$$

Hence $V(\hat{\bar{x}}) >, =, < V(\bar{x})$ according as

$$C(N_iX_i) <, =, > -[(M-1)/(n-1)](1-n/N).$$

(ii) *When $2C(N_iX_i) = C(N_iN_i)$,* i.e., *for ordinary cluster sampling the ratio estimate and unbiased estimate have the same variance.* In such situations

$$V(\hat{\bar{x}}) = V(\bar{x}) - [(M-m)/mn][1-(n/N)]R^2,$$

and thus $V(\bar{x}) > V(\hat{\bar{x}})$ and $V(\hat{\bar{x}})$ decreases sharply as R increases.

4. Selection of post clusters with probabilities proportional to size. The problem of selecting with varying probabilities is well known and the general theory has

been discussed by many, and as our task is only an application of these general theories to a special stochastic model, hence we shall not take the trouble of exploring all branches of the general theories but will work with only one case, namely drawing m clusters with probabilities proportional to the post cluster sizes and sampling being done with replacement. Hence the clusters will be selected with probabilities

$$p_i = \sum_{j=1}^{N_i} \epsilon_{ij}/n, \qquad i = 1, 2, \cdots, M.$$

Unbiased estimate. Suppose $d_i = \sum_{j=1}^{N_i} \epsilon_{ij} X_{ij}$ and let $Z_i = d_i/p_i n$.

LEMMA[4] 5. *$\bar{Z} = \sum_{i=1}^{m} Z_i/m$ is an unbiased estimate of $\bar{X}$.*

PROOF. It is easy to see that the conditional expectation of Z_i given the p_i's is $E(Z_i/p) = \sum_{i=1}^{M} d_i/n$ which is the same for all i, hence

$$E(\bar{Z}) = \sum_{i=1}^{M} E(d_i)/n = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij} E(\epsilon_{ij})/n = X/N = \bar{X},$$

which completes the proof.

LEMMA 6. *The variance of $\bar{Z}$ is given by*

$$V(\bar{Z}) = [mn(N-1)]^{-1} \Big[(N-n) \sum_{i=1}^{M} \sigma_{Wi}^2 + (n-1)N\sigma_B^2 + (m-1)(N-n)\sigma^2 + (N-n) \sum_{i=1}^{M} (\bar{X}_i^2 - \bar{X}^2/M) + n(m-1)(N-1)\bar{X}^2 \Big] + o(N_i)^{-1},$$

where

$\sigma_{Wi}^2 = (N_i)^{-1} \sum_{j=1}^{N_i} X_{ij}^2 - \bar{X}^2$ = within variance of the ith cluster.

$\sigma_B^2 = N^{-1} \sum_{i=1}^{M} (\bar{X}_i - \bar{X})^2$ = between variance of the means of the clusters.

The proof of the above lemma is given in the appendix.

In the special case when the within variances are the same, say σ_W^2, in each cluster the expression for the variance becomes

$$V(\bar{Z}) = [mn(N-1)]^{-1} \Big[(N-n)M\sigma_W^2 + (n-1)N\sigma_B^2 + (m-1)(N-n)\sigma^2 + (N-n) \sum_{i=1}^{M} (\bar{X}_i^2 - \bar{X}^2/M) + n(m-1)(N-1)\bar{X}^2 \Big]. \tag{9}$$

Though (9) is not a very neat expression, yet it indicates that for sampling with probability proportional to size PCS σ_W^2, σ_B^2, σ^2, $\bar{X}_i^2$, $\bar{X}^2$, all contribute to increase the variance of $\bar{Z}$.

[4] In actual realization some p_i may be zero, in such situations that particular cluster for which p_i is zero is omitted while drawing samples from the post clusters.

5. Optimum design for PCS. It is obvious that in PCS there are two parameters n and m which can be chosen in an arbitrary manner except for the restriction $n \leqq N$, $m \leqq M$. Here the problem of selecting these parameters in an optimum manner shall be discussed. The cost plays an important role in all sample surveys and the survey has to be designed within a fixed budget. In PCS the number of elements ultimately to be surveyed is a random variable, hence the restriction that shall be imposed is that the expected cost should be equal to the budget C_0. The cost function can be assumed to be linear, say, $C = C_1 n + C_2 \sum_{i=1}^{M} \sum_{j=1}^{N_i} \phi_i \epsilon_{ij}$, where

C_1 = Cost per element of the random sample,

C_2 = Cost per element finally selected through the cluster sample. Hence the expected cost is given by $E(C) = C_1 n + C_2 mn$,

If expected cost is fixed at C_0, then the minimum variance unbiased estimate is obtained with

$$(10) \quad m = \frac{\{(\sigma^2 + \bar{X}^2)(NC_1 - C_0)MN + M^2(C_0 - C_1)V_c\}^{\frac{1}{2}}}{(MV_c - \bar{X}^2N^2)^{\frac{1}{2}}C_2^{\frac{1}{2}}},$$

$$(11) \quad n = \frac{C_0(MV_c - \bar{X}^2N^2)^{\frac{1}{2}}}{C_1(MV_c - \bar{X}^2N^2)^{\frac{1}{2}} + C_2^{\frac{1}{2}}\{(\sigma^2 + \bar{X}^2)(NC_1 - C_0)MN + M^2(C_0 - C_1)V_c\}^{\frac{1}{2}}}.$$

These solutions are obtained by routine minimization techniques using Lagrange multipliers. In the simplification $N - 1$ and $M - 1$ have been approximated by N and M respectively.

6. Appendix. Variance of $\bar{x}$.

$$V(\bar{x}) = (M^2/m^2n^2)[\sum_i \sum_j X_{ij}^2 V(\phi_i \cdot \epsilon_{ij}) + \sum_i \sum_{j \neq l} X_{ij}X_{il} \operatorname{Cov}(\phi_i\epsilon_{ij}, \phi_i\epsilon_{il}) + \sum_{i \neq k} \sum_j \sum_l X_{ij}X_{kl} \operatorname{Cov}(\phi_i\epsilon_{ij}, \phi_k\epsilon_{kl})].$$

Using Lemmas 2, 3, and 4 we get

$$V(\bar{x}) = \frac{M^2}{m^2n^2}\left[\sum_i \sum_j X_{ij}^2 \left\{\frac{n}{N} \cdot \frac{m(M-m)}{M^2} + \frac{m^2}{M^2} \cdot \frac{n(N-n)}{N^2}\right\} + \sum_i \sum_{j \neq l} X_{ij}X_{il} \left\{\frac{m(M-m)}{M^2} \cdot \frac{n(n-1)}{N(N-1)} - \frac{m^2}{M^2} \cdot \frac{n(N-n)}{N^2(N-1)}\right\} + \sum_{i \neq k} \sum_{j,l} X_{ij}X_{kl} \left\{\frac{n(n-1)}{N(N-1)} \frac{m(M-m)}{M^2(M-1)} - \frac{m^2}{M^2} \frac{n(N-n)}{N^2(N-1)}\right\}\right].$$

The quantities in the second bracket can be taken outside the summation sign and then making use of the facts

$$\frac{1}{N_i}\sum_j X_{ij} = \bar{X}_i = X_i/N_i$$

$$\sum_i \sum_{j \neq l} X_{ij}X_{il} = \sum_i (\sum_j X_{ij})^2 - \sum_i \sum_j X_{ij}^2 = \sum_i X_i^2 - \sum_i \sum_j X_{ij}^2$$

$$\sum_{i\neq k}\sum_{j,l} X_{ij}X_{kl} = (\sum_i \sum_j X_{ij})^2 - \sum_i (\sum_j X_{ij})^2 = N^2\bar{X}^2 - \sum_i X_i^2,$$

and using the definitions of σ^2 and V_c as given in Section 3 and on simplifying a little we get Equation (1). To obtain (2) we add and subtract $V(R_c)$ on the right-hand side of (1) and simplify a little.

Calculation of bias in $\hat{x}$. $\hat{x}$ can be written in the form

$$\hat{x} = R \cdot \frac{\sum_i \sum_j \phi_i \epsilon_{ij} X_{ij}/(mn/MN)X}{\sum_i \sum_j \phi_i \epsilon_{ij}/(mn/MN)N}$$

$$= R\left[1 + \frac{\sum_i \sum_j (\phi_i \epsilon_{ij} - mn/MN)X_{ij}}{(mn/MN)X}\right]\left[1 + \frac{\sum_i \sum_j (\phi_i \epsilon_{ij} - mn/MN)}{(mn/MN)N}\right]^{-1}.$$

The second factor can be expanded in Taylor series and then multiplying term by term and neglecting terms of smaller order than $o(1/N^2)$ we have

$$\hat{x} = R\left[1 + \frac{\sum_i \sum_j (\phi_i \epsilon_{ij} - mn/MN)X_{ij}}{(mn/MN)X} - \frac{\sum\sum (\phi_i \epsilon_{ij} - mn/MN)}{(mn/MN)N} + \frac{\{\sum_i \sum_j (\phi_i \epsilon_{ij} - mn/MN)\}^2}{(mn/MN)^2N^2} - \frac{\{\sum_i \sum_j (\phi_i \epsilon_{ij} - mn/MN)X_{ij}\}\{\sum_i \sum_j (\phi_i \epsilon_{ij} - mn/MN)\}}{(mn/MN)^2NX} + o(1/N^2)\right].$$

A first approximation to the bias is obtained by taking expectation of the above expression. Hence

$$E(\hat{x}) - R \approx R\left[\frac{V(\sum_i \sum_j \phi_i \epsilon_{ij})}{(mn/MN)^2N^2} - \frac{\mathrm{Cov}\,(\sum_i \sum_j \phi_i \epsilon_{ij} X_{ij}, \sum_i \sum_j \phi_i \epsilon_{ij})}{(mn/MN)^2NX}\right].$$

We have already calculated $V(\sum_i \sum_j \phi_i \epsilon_{ij} X_{ij})$ and if we substitute in that expression $X_{ij} \equiv 1$ and note the following $\sum_i \sum_j X_{ij}^2 = \sum_i \sum_j 1 = N$, for fixed i, $\sum_{j\neq k} 1 = N_i(N_i - 1)$, $\sum_{i\neq k}\sum_{j,l} 1 = \sum_{i\neq k} N_iN_k$ then

$$V(\sum\sum \phi_i \epsilon_{ij}) = \frac{mn}{M^2N^2}\left[(MN - mn)N + \frac{MNn - Nmn - MN + mn}{N-1}\sum_{i=1}^{M} N_i(N_i - 1) + \frac{MN - mn + Mmn + Nmn - MNn - MNm}{(M-1)(N-1)}\sum_{i\neq k} N_iN_k\right].$$

Now making use of the fact that $\sum_{i\neq k} N_iN_k = (\sum_i N_i)^2 - \sum_i N_i^2 = N^2 - \sum_i N_i^2$ and $\sigma^2(N_i) = \sum_i N_i^2/M - N^2/M^2$ and simplifying a little we have

$$V(\sum\sum \phi_i\epsilon_{ij}) = [mn(M-m)/M^2N^2(M-1)(N-1)] \cdot [M^2N(n-1)\sigma^2(N_i) + N^2(N-n)(M-1)].$$

Again $\mathrm{Cov}\,(\sum_i \sum_j \phi_i\epsilon_{ij}X_{ij}, \sum_i \sum_j \phi_i\epsilon_{ij})$ can be obtained from $V(\sum_i \sum_j \phi_i\epsilon_{ij}X_{ij}) = \mathrm{Cov}\,(\sum_i \sum_j \phi_i\epsilon_{ij}X_{ij}, \sum_i \sum_j \phi_i\epsilon_{ij}X_{ij})$ by substitut-

ing unity for one set of X_{ij}'s. Thus

$$\text{Cov}\left(\sum_i \sum_j \phi_i \epsilon_{ij} X_{ij}, \sum_i \sum_j \phi_i \epsilon_{ij}\right)$$

$$= \frac{mn}{M^2N^2}\left[N(MN - mn)\bar{X} + \frac{Nn(M-m) - MN + mn}{N-1}\sum_i (N_i - 1)\sum_j X_{ij} + \frac{MN - mn + Mmn + Nmn - MNn - MNm}{(M-1)(N-1)}\sum_{i\neq k} N_i \sum_j X_{kj}\right].$$

Noting the fact that $\sum_{i\neq k} N_i \sum_j X_{kj} = \sum_{i\neq k} N_i X_k = NX - \sum_i N_i X_i = N^2\bar{X} - \sum_i N_i X_i$ and using the same type of simplification as in calculating $V(\sum_i \sum_j \phi_i \epsilon_{ij})$ we have

$$\text{Cov}\left(\sum_i \sum_j \phi_i \epsilon_{ij} X_{ij}, \sum_i \sum_j \phi_i \epsilon_{ij}\right) = [mn(M-m)/M^2N^2(M-1)(N-1)]$$

$$[M^2N(n-1)\,\text{Cov}(N_i, X_i) + N^2\bar{X}(N-n)(M-1)].$$

Hence

$$\text{Bias} = E(\hat{x}) - R = \frac{R(M-m)}{mn(M-1)(N-1)} \cdot \left[\frac{M^2(n-1)}{N}\sigma^2(N_i) + (N-n)(M-1) - \frac{M^2(n-1)}{N\bar{X}}\,\text{Cov}(N_i, X_i) - (N-n)(M-1)\right]$$

$$= R\,\frac{(M-m)(n-1)N}{mn(M-1)(N-1)}\,[C(N_i N_i) - C(N_i X_i)].$$

Calculation of the variance of $\bar{Z}$.

$$V(\bar{z}) = E_{R_1} V(\bar{Z}/R_1) + V_{R_i} E(\bar{Z}/R_i)$$

$$= E_{R_1}\frac{1}{m}\left[\sum_{i=1}^{M} p_i Z_i^2 - E^2(\bar{Z}/R_1) + \frac{1}{n^2} V_{R_1}\left(\sum_{i=1}^{M}\sum_{j=1}^{N_i} \epsilon_{ij} X_{ij}\right)\right.$$

$$= \frac{1}{m}\sum_i E_{R_1}\frac{d_i^2}{p_i n^2} - \frac{1}{n^2 m} E_{R_1}(\sum_i d_i)^2 + \frac{1}{n^2} V_{R_1}(\sum_i \sum_j \epsilon_{ij} X_{ij})$$

$$= \frac{1}{n^2}\left[\frac{n}{m}\sum_i E_{R_1}\left(\frac{(\sum_j \epsilon_{ij} X_{ij})^2}{\sum_j \epsilon_{ij}}\right) + \left(1 - \frac{1}{m}\right) E_{R_1}(\sum_i \sum_j \epsilon_{ij} X_{ij})^2 - (\sum_i \sum_j X_{ij} E_{R_1}(\epsilon_{ij}))^2\right].$$

Now

$$(\sum_i \sum_j X_{ij} E_{R_1}(\epsilon_{ij}))^2 = [\sum_i \sum_j X_{ij}(n/N)]^2 = n^2\bar{X}^2.$$

$$E_{R_1}(\sum_i \sum_j \epsilon_{ij} X_{ij})^2 = \frac{n(n-1)N}{(N-1)}\bar{X}^2 + \frac{n(N-n)}{N(N-1)}\sum_i \sum_j X_{ij}^2.$$

$$E_{R_1}\left(\frac{(\sum_j \epsilon_{ij} X_{ij})^2}{\sum_j \epsilon_{ij}}\right) = \frac{1}{N-1}\left[(N-n)\frac{1}{N_i}\sum_j X_{ij}^2 + (n-1)N_i\bar{X}_i^2\right] + o\left(\frac{1}{N_i}\right).$$

Substituting the above in $V(\bar{Z})$ and simplifying we get (9).

7. Acknowledgment. The author is grateful to Professor J. L. Hodges, Jr. for his constant guidance and help in the course of the entire work. The author wishes to thank Dr. T. Dalenius for suggesting the problem and making some helpful comments on the manuscript. The author wishes to thank the referee for some helpful comments.

REFERENCES

[1] DALENIUS, T. (1957). *Sampling in Sweden.* Almqvist and Wiksell, Stockholm.
[2] DAVID, F. N. (1938). Limiting distributions connected with certain methods of sampling human populations. *Statist. Res. Mem. London* **2** 69–90.
[3] GOODMAN, L. (1960). On the exact variance of products. *J. Amer. Statist. Assoc.* **55** 708–713.
[4] LAHIRI, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Internat. Statist. Inst.* **33** No. 2 133–140.
[5] NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.* **33** 101–116.
[6] QUENOUILLE, M. H. (1958). *Fundamentals of Statistical Reasoning.* Griffin, London.