ESTIMATORS IN MULTIPLE FRAME SURVEYS

Richard E. Lund

Iowa State University and
Centro de Estadistica y Cálculo, Chapingo, México

Estimators appror riate for multiple frame surveys were proposea by Hartley [3]. This paper suggests an alteration in these estimators consisting of basing the weights associated with the sample from each frame upon the actual sample sizes obtained. The resulting estimators have equal or greater efficiency. Complexity is reduced in both the estimator and sample allocation determination.

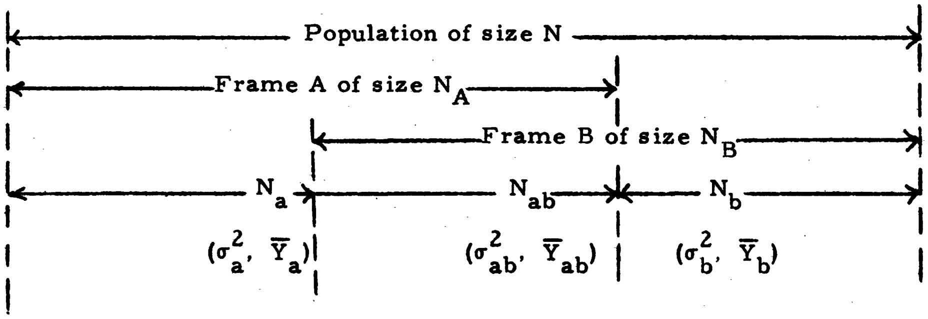
Introduction

A sampling frame or list is the keystone around which a sampling process is constructed. But often a single list corresponding to all elements in the desired population is not available. Consequently, two or more lists are frequently used to construct a frame of satisfactory coverage. In other situations a single suitable, but relatively costly, frame is available, however efficiency considerations suggest the joint use of another less costly incomplete frame.

Commonly used sampling estimators require the elimination of duplicated elements from the frame. However, such an elimination can be impractical. Hartley [3] proposed estimators and allocation formula suitable for use with two overlapping frames. This paper suggests alterations in these estimators which improve efficiency and decrease complexity.

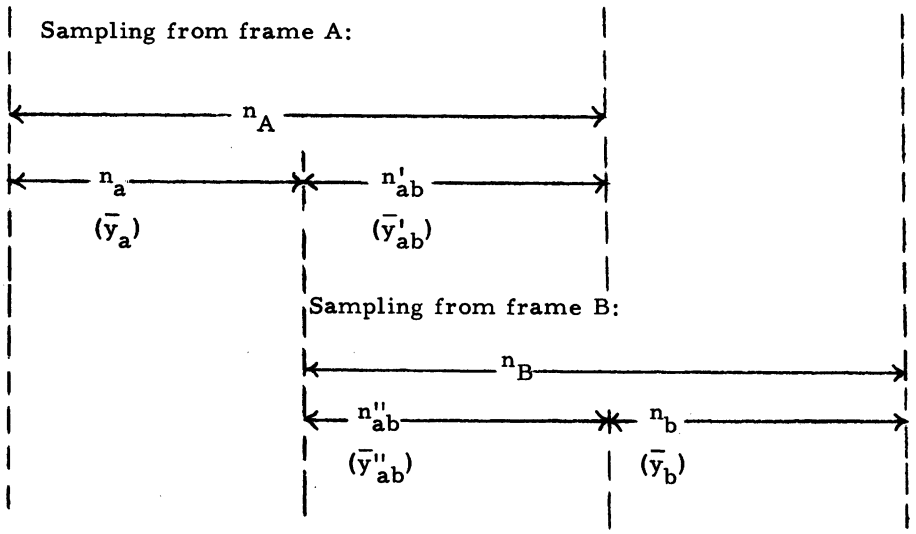
Notation and Nomenclature

The survey objective is considered to be the estimation of the total (ΣY_i) of characteristic "Y" for a population containing N elements. Complete coverage of the desired population is provided by two overlapping frames A and B of sizes N_A and N_B. The population can be separated into three domains: (a) the non-duplicated elements associated with frame A, (ab) the elements duplicated in both A and B, and (b) the non-duplicated elements in B. Using N_a, N_{ab} and N_b to represent the size of each domain, $\overline{Y}_a$, $\overline{Y}_{ab}$ and $\overline{Y}_b$, and σ_a^2, σ_{ab}^2 and σ_b^2 to represent the population means and unit variances, the basis can be presented schematically as:


Population of size N
Frame A of size N_A
Frame B of size N_B
N_a
N_{ab}
N_b
$(\sigma_a^2, \overline{Y}_a)$
$(\sigma_{ab}^2, \overline{Y}_{ab})$
$(\sigma_b^2, \overline{Y}_b)$

Several expressions are simplified by referring to the relative size of the overlap with respect to frame size $(\alpha = N_{ab}/N_a$ and $\beta = N_{ab}/N_B)$.

Random samples of n_A and n_B elements are selected independently from the two frames. The division of the sample for any frame between the two domains (unduplicated and duplicated elements) is not considered to be subject to control by the sampler. Symbols n_a and n'_{ab} refer to the sample sizes resulting from the random division of n_A. Sizes n''_{ab} and n_b are associated similarly with the sample from frame B. Defining the sample means $\bar{y}_a$, $\bar{y}'_{ab}$, $\bar{y}''_{ab}$ and $\bar{y}_b$, we have:



Having this basis, attention can be turned to two principal cases: first, the case of domain sizes N_a, N_{ab} and N_b known and second, the case of N_a, N_{ab} and N_b unknown but N_A and N_B known with estimates of α and β available for sample allocation.

Case of N_a, N_{ab} and N_b Known

Hartley suggested the unbiased estimator of the population total

$$\hat{Y}_H = N_a\bar{y}_a + N_{ab}p\bar{y}'_{ab} + N_{ab}(1-p)\bar{y}''_{ab} + N_b\bar{y}_b \quad (1)$$

where $o \leq p \leq 1$. The variance of this estimator is approximately

$$\begin{aligned} \mathrm{Var}(\hat{Y}_H) &\doteq \frac{N_A^2}{n_A}\,[(1-\alpha)\sigma_a^2 + \alpha p^2\sigma_{ab}^2] \\ &+ \frac{N_B^2}{n_B}\,[(1-\beta)\sigma_b^2 + \beta(1-p)^2\sigma_{ab}^2]\,. \quad (2) \end{aligned}$$

Finite population corrections have been ignored.

Sampling costs can be expressed by the linear function

$$\text{Total Cost} = n_Ac_A + n_Bc_B \quad (3)$$

where c_A and c_B define unit costs of sampling from each frame respectively. The problem of optimizing the sample allocation among the two frames and finding the optimum value for p consists of minimizing (2) as a function of n_A, n_B and p subject to restriction (3). Hartley expressed the value for p as a bi-quadratic while additional formulas were given for n_A and n_B. The solution for p, however, has the simple expression

$$p_o = \frac{\alpha n_A}{\alpha n_A + \beta n_B} \quad (4)$$

Thus, the optimum value for p is the ratio of the expected value of n'_{ab} with respect to the expected value of $n'_{ab} + n''_{ab}$.

Hartley's procedure does not consider the actual division achieved (at random) of the n_A and n_B elements among the domains. Thus, one may ask whether a gain is achieved by making p a function of n'_{ab} and n''_{ab}.

To reach a solution, the variance of (1) can be taken in two steps by use of the well known theorem expressed in symbols of the current problem

$$\begin{aligned} \mathrm{Var}(\hat{Y}) &= E[\mathrm{Var}(\hat{Y}\,|\,n'_{ab},\, n''_{ab})] \\ &+ \mathrm{Var}[E(\hat{Y}\,|\,n'_{ab},\, n''_{ab})] \quad (5) \end{aligned}$$

where the condition, $|n'_{ab}, n''_{ab})$, represents the actual sample counts achieved. The variable within the second term, $E[(\hat{Y}\,|\,n'_{ab}, n''_{ab})]$, equals the population total for any value of p. Thus, its variance equals zero.

Disregarding finite population corrections,

$$\begin{aligned} \mathrm{Var}(\hat{Y}\,|\,n'_{ab},\, n''_{ab}) &= \frac{N_a^2}{n_a}\sigma_a^2 + p^2\,\frac{N_{ab}^2}{n'_{ab}}\,\sigma_{ab}^2 \\ &+ (1-p)^2\,\frac{N_{ab}^2}{n''_{ab}}\sigma_{ab}^2 + \frac{N_b^2}{n_b}\,\sigma_b^2. \quad (6) \end{aligned}$$

Minimization of (6) as a function of p gives the solution

$$p_o = \frac{n'_{ab}}{n'_{ab} + n''_{ab}} \quad . \quad (7)$$

The variance for the estimator with this value for p can be found by substituting (7) into (6) and obtaining the expected value. The estimator and its approximate variance are

$$\hat{Y}_L = N_a\bar{y}_a + N_{ab}\bar{y}_{ab} + N_b\bar{y}_b \qquad (8)$$

where $$\bar{y}_{ab} = \frac{n'_{ab}\bar{y}'_{ab} + n''_{ab}\bar{y}''_{ab}}{n'_{ab} + n''_{ab}}$$

and

$$\mathrm{Var}(\hat{Y}_L) \doteq \frac{N_A^2}{n_A}(1-\alpha)\sigma_a^2 + \frac{N_A N_B \alpha\beta}{\alpha n_A + \beta n_B}\sigma_{ab}^2 + \frac{N_B^2}{n_B}(1-\beta)\sigma_b^2 \ . \qquad (9)$$

The order of the approximation in (9) is the same as for (2).

While it can be proven that estimator (8) is always equal or greater in efficiency than (1), this increase in efficiency is not reflected in variance approximation (9). Substitution of (4) into (2) provides an expression identical to (9) which indicates that both estimators are equal in efficiency to the order of the approximation.

The departure of variance approximation (9) from the true variance of $\hat{Y}_L$ was estimated by examining terms through the second order of a Taylor's expansion of the random variables $1/n_a$, $1/(n'_{ab} + n''_{ab})$ and $1/n_b$ around the values $1/\alpha n_A$, $1/(\alpha n_A + \beta n_B)$ and $1/\beta n_B$ respectively. The first term of (9) could be corrected by multiplying by $[1 + \alpha/(1-\alpha)n_A]$ and the third term by $[1 + \beta/(1-\beta)n_B]$. The correction for the second term is $[1 + \delta/(\alpha n_A + \beta n_B)]$ where δ is a weighted average of $(1-\alpha)$ and $(1-\beta)$, the weights being αn_A and βn_B. As noted earlier, approximation (9) is also appropriate for the variance of estimator (1) with p expressed by (4), but δ in the second term correction becomes the sum of $(1-\alpha)$ and $(1-\beta)$. No change occurs in the correction for the first and third terms. Thus, it is seen that variance approximation (9) (or 2) is reasonably accurate for all but very small samples and in addition, the gain in efficiency by use expression (7) for p instead of (4) is negligible except for extremely small samples.

The general solution of the allocation problem, as found by minimizing (9) as a function of n_A and n_B subject to cost equation (3) can be expressed by the iterative system

$$r_1 = \sqrt{\frac{c_B}{c_A}\left(\frac{\beta}{\alpha}\right)}$$

$$r_{i+1}^2 = \frac{c_B}{c_A}\left(\frac{\beta}{\alpha}\right)^2 \times \frac{(r_i+\frac{\beta}{\alpha})^2(1-\alpha)\sigma_a^2 + r_i^2\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2(1-\beta)\sigma_b^2 + (\frac{\beta}{\alpha})^2\beta\sigma_{ab}^2} \qquad (10)$$

where $r = n_A/n_B$. Practice with the system for several values of the parameters indicated that few iterations are required in most cases.

To determine the sensitivity of the estimator to deviations from optimum allocation, the optimum value for r and the corresponding variance were computed for a wide range of values of the parameters. A comparison was made to the variances corresponding to deviations of ten percent in both directions from the optimum (that is, $0.90r_o$ and $1.10r_o$). The variance was increased by more than one percent in very few instances by the deviation from optimum allocation.

The case of complete coverage by a relatively costly frame merits some additional consideration. Defining A as the complete frame, it is clear that $N_b = 0$ and $\beta = 1$ which enable a simple graphical presentation of the optimum allocation. Figure 1 displays the solutions for four relative cost levels and three variance ratios.

Case of N_a, N_{ab} and N_b Unknown

Estimators (8) can serve as the starting point for the case of unknown N_a, N_{ab} and N_b. However, it is necessary to insert estimates for these sizes by use of sample data and known N_A and N_B. The expressions $N_A(n_a/n_A)$ and $N_B(n_b/n_B)$ are unbiased estimators of N_a and N_b, respectively. Two unbiased estimators of N_{ab} are available: $N_A(n'_{ab}/n_A)$ and $N_B(n''_{ab}/n_B)$. Using p and (1-p) as undetermined weights for the two estimates of N_{ab} and substituting these expressions in (8), an unbiased estimator for the population total is

$$\dot{Y} = \frac{N_A}{n_A}n_a\bar{y}_a + [\frac{N_A}{n_A}n'_{ab}p + \frac{N_B}{n_B}n''_{ab}(1-p)]\bar{y}_{ab} + \frac{N_B}{n_B}n_b\bar{y}_b \qquad (11)$$

where as before $\bar{y}_{ab}$ is the sample mean of all elements selected from the domain of duplicated elements.

Use of theorem (5) provides the approximate variance

$$\mathrm{Var}(\dot{Y}) \doteq \frac{N_A^2}{n_A}(1-\alpha)\sigma_a^2 + \frac{N_A N_B \alpha\beta}{\alpha n_A + \beta n_B}\sigma_{ab}^2$$

$$+ \frac{N_B^2}{n_B}(1-\beta)\sigma_b^2$$

$$+ \frac{N_A^2(1-\alpha)\alpha}{n_A}[\bar{Y}_a - p\,\bar{Y}_{ab}]^2$$

$$+ \frac{N_B^2(1-\beta)\beta}{n_B}[\bar{Y}_b-(1-p)\,\bar{Y}_{ab}]^2 \,. \qquad (12)$$

The final two terms represent the increase in variance due to not knowing N_a, N_{ab} and N_b. These terms make a significant contribution unless either the overlap is nearly complete or is relatively small.

The degree of the approximation in variance (12) is not the same as noted earlier for (9) or (2). All terms are exact except for the second. Examination of terms through the second order of a Taylor's expansion of the second term in $\mathrm{Var}(\dot{Y}\,|\,n'_{ab}, n''_{ab})$ suggests a multiplicative correction equal to or less than $[1+(1-\alpha)/\alpha n_A + (1-\beta)/\beta n_B]$. Thus, the approximation is reasonably accurate for all practical uses.

Minimization of (12) as a function of p, n_A and n_B subject to cost equation (3) specifies

$$p_o = \frac{\frac{N_A(1-\alpha)}{n_A}\bar{Y}_a + \frac{N_B(1-\beta)}{n_B}(\bar{Y}_{ab}-\bar{Y}_b)}{[\frac{N_A(1-\alpha)}{n_A} + \frac{N_B(1-\beta)}{n_B}]\,\bar{Y}_{ab}} \qquad (13)$$

and provides the sample allocation among the two frames. The sample allocation is expressed again as an iterative system

$$r_1 = \sqrt{\frac{c_B}{c_A}(\frac{\beta}{\alpha})}$$

$$r_{i+1}^2 = \frac{c_B}{c_A}(\frac{\beta}{\alpha})^2\left[(1-\alpha)\sigma_a^2 + \frac{r_i^2\,\alpha\,\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2}\right.$$

$$\left.+ \frac{r_i^2\alpha(1-\alpha)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}\right]$$

$$X\left[(1-\beta)\sigma_b^2 + \frac{(\frac{\beta}{\alpha})^2\beta\,\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2}\right.$$

$$\left.+ \frac{[\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2\,\beta(1-\beta)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}\right]^{-1} \,. \qquad (14)$$

Use of the system with various values of the parameters indicates that normally few iterations are required.

It is apparent that the sample allocation is based on preliminary estimates of the unit variances and population means as well as upon preliminary estimates of N_a, N_{ab} and N_b as expressed in α and β. A numerical investigation similar to that discussed for the case of N_a, N_{ab} and N_b known, indicates the efficiency of the estimator is rather insensitive to moderate departures from optimum allocation.

To determine the value of p, one may,of course,utilize the estimates of the parameters used in determining the sample allocation. However, some gain in efficiency would probably be achieved by using the sample data. An estimator of optimum p from the sample data is

$$\hat{p} = \frac{\frac{N_A n_a}{n_A^2}\bar{y}_a + \frac{N_B n_b}{n_B^2}(\bar{y}_{ab}-\bar{y}_b)}{\left(\frac{N_A n_a}{n_A^2} + \frac{N_B n_b}{n_B^2}\right)\bar{y}_{ab}} \,. \qquad (15)$$

Of course, use of (15) disturbs the unbiasedness of estimator (11) for p is now a function of n'_{ab} and n''_{ab}. However, the degree of bias is considered to be negligible. The expected value of estimator (11) with p given by (15) was approximated by use of a Taylor's expansion in which terms through the second order were retained. This approximation suggested the second term will be biased downward by the multiplicative factor $[1-\delta]$ where δ is the weighted mean of $(1/n_A)$ and $(1/n_B)$, the weights being $N_A(1-\alpha)/n_A$ and $N_B(1-\beta)/n_B$.

Hartley suggested an estimator to handle the current case, but his weight variable p was used to weight two individual estimates of the total over the duplicated elements $(N_{ab}\,\bar{Y}_{ab})$. The expression for the optimum value for such a p includes the parameters σ_a^2, σ_{ab}^2 and σ_b^2 . Thus, the estimator suggested herein is somewhat simplier. But in addition it can be proven that estimator (11) always has equal or greater efficiency than the estimator suggested by Hartley.

The reduction in variance can be quite important. In the not unusual case of equal per unit variances, equal means among the various strata, coefficient of variation of .10, $\alpha = \beta = .80$, and relative costs of $c_A/c_B = 4$, the reduction in variance by use of (11) was about one-fourth. Cases requiring more extreme sample allocation between the two frames, by reason for example of greater cost differences, show a greater gain in efficiency by use of (11).

Figure 2 presents the optimum allocation for the special case of $\alpha = \beta$, equal unit variances and approximately equal domain means. Four relative cost levels and three values for the coefficient of variation are considered.

Numerical Example

It may be assumed that a survey is to be conducted to measure characteristics of daily milk consumption in Mexico City. The sampling unit selected is the household. Among the various sampling frames available, it is decided to use the following two:

1. Telephone directory
2. Housing registration records maintained by the Secretary of Housing.

Preliminary investigations indicate the second frame covers the desired population satisfactorily. While the first frame provides inadequate coverage, costs of using it are substantially less because the data can be collected by telephone. It is assumed to be more economical to collect all data by personal interview for households selected from the second frame by reason of difficulties in matching housing unit records to telephone numbers.

It is estimated that the telephone directory provides 40 percent coverage. Per unit variances are assumed to be equal in both the duplicated parts of the population. Applying this information to Figure 1 suggests the sample allocation of 50 percent to each frame. If the relative cost ratio of the housing records with respect to the telephone directory were only 2, the optimum allocation would be to assign only 20 percent to the lower cost frame.

As an example of different conditions, suppose that a similar survey is planned for a U.S. city. The investigator decides to use two frames similar to those suggested earlier. However, in this case preliminary investigations show that each frame covers about 90 percent of the desired population while complete coverage is provided by using both frames. The 90 percent value is only a crude estimate for survey planning purposes and the investigator wishes to treat the domain sizes as unknown (N_a, N_{ab} and N_b). Supposing that the per unit costs of using the telephone directory are only one-half those of using a housing unit list and the variances and means for the various domains are equal approximately with a coefficient of variation of 0.5, the optimum allocation indicated by Figure 2 is to assign 75 percent of the sample to the lower cost frame.

It is clear that many practical situations cannot be handled by Figures 1 and 2. The optimum allocation for these can be determined by iterative systems (10) or (14).

References

[1] Cochran, Robert S. (1964). "Multiple Frame Sample Surveys." Proceedings of Social Science Section of American Statistical Association meetings, Chicago, Illinois.

[2] Cochran, Robert S. (1967). "The Estimation of Domain Sizes When Sampling Frames Are Interlocking." Mimeographed paper at the American Statistical Association meetings, Social Science Section, Washington, D.C.

[3] Hartley, H.O. (1962). "Multiple Frame Surveys." Proceedings of the Social Science Section of American Statistical Association meetings, Minneapolis, Minnesota.

[4] Steinberg, J. (1965). "A Multiple Frame Survey for Rare Population Elements." Proceedings of the Social Science Section of American Statistical Association meetings.

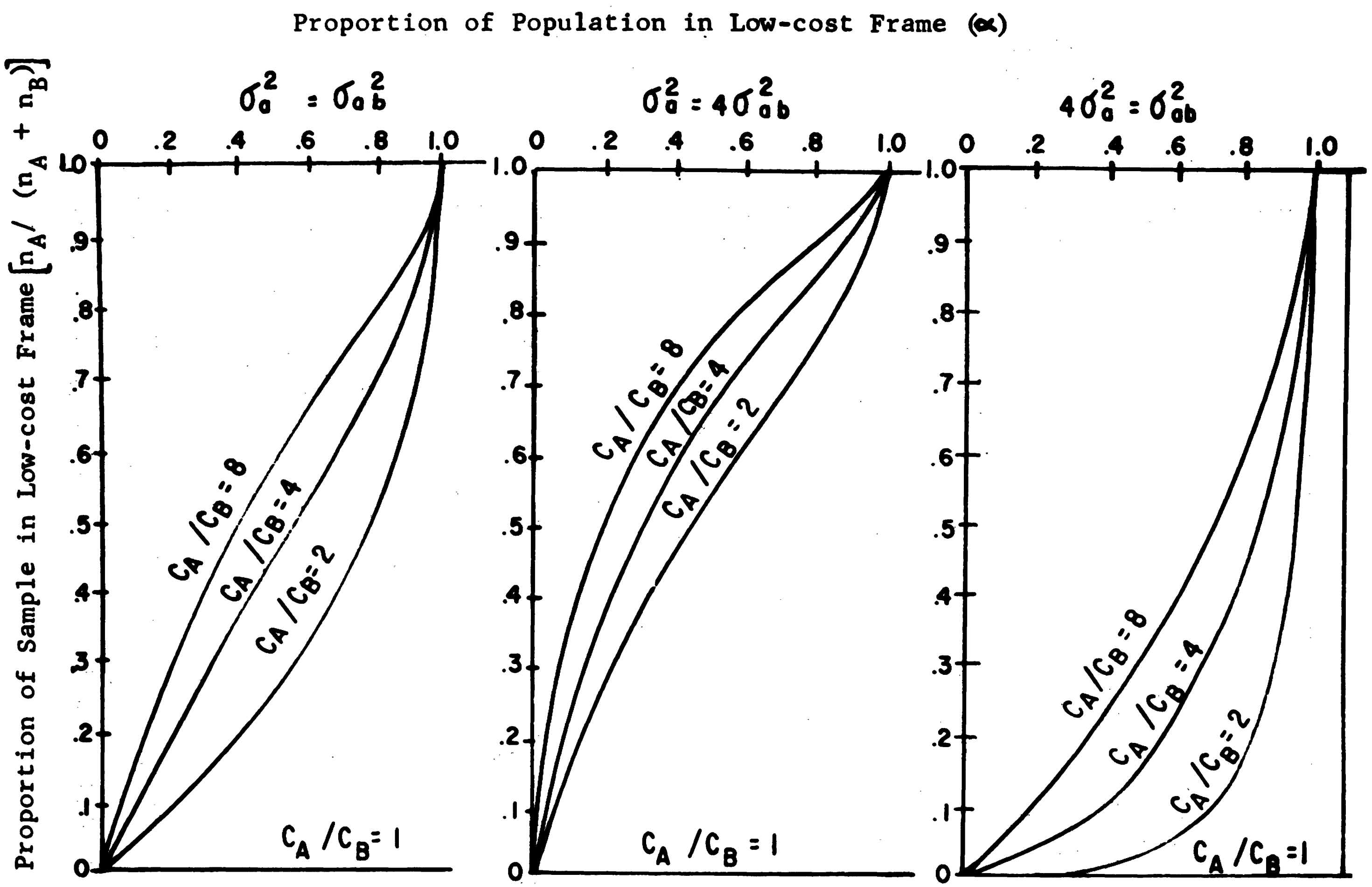


Figure 1

Optimum Allocation for Special Case of Complete Coverage
by Costly Frame

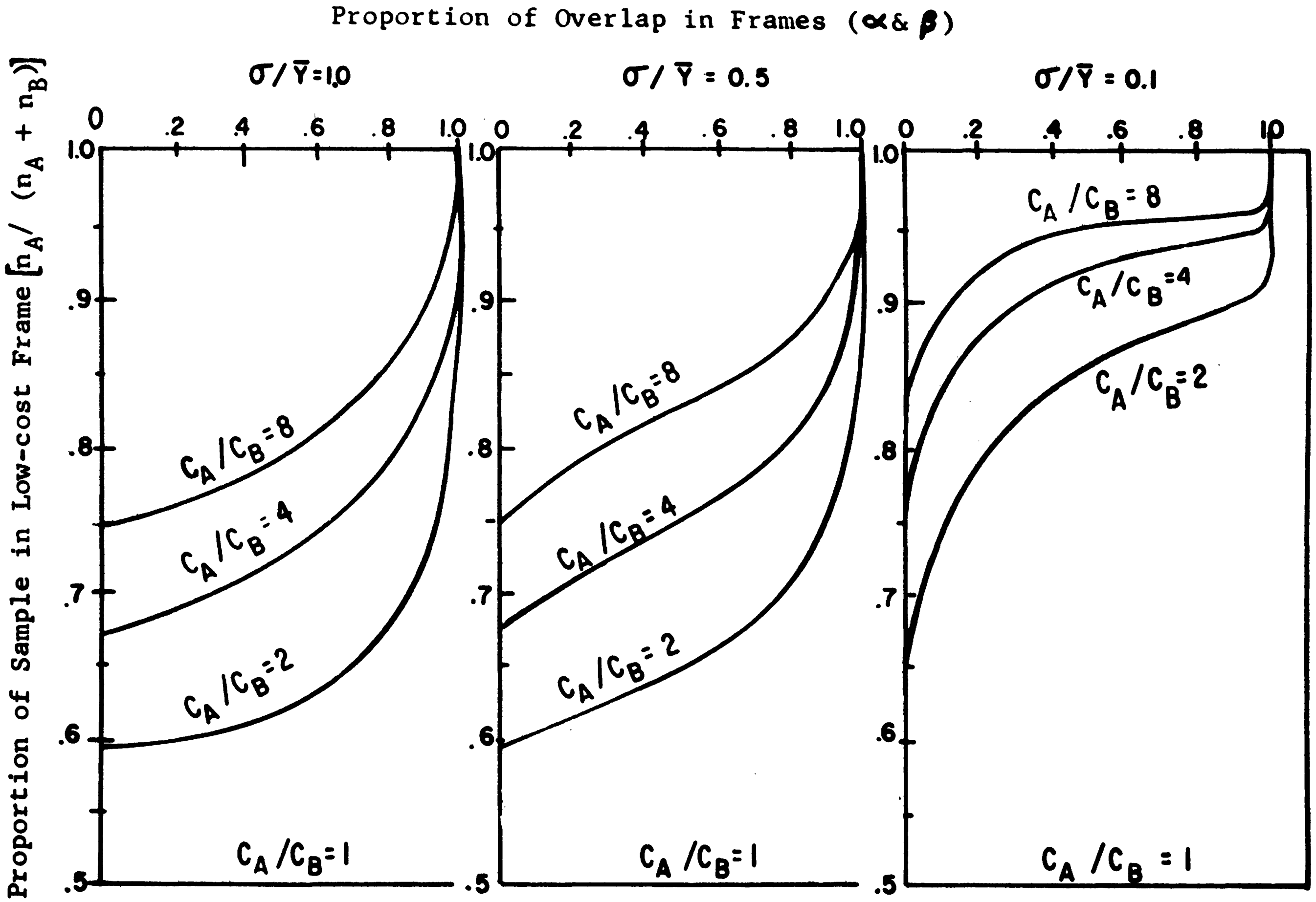


Figure 2

Optimum Allocation for Special Case of Equal Frame Sizes, Equal Variances and Means (Approx.), and Unknown Domain Sizes (N_a, N_{ab}, N_b)