MULTIPLE FRAME SURVEYS

H. O. Hartley

1. Concepts.

The situations encompassed by the term 'multiple frame surveys' may be described as follows: In sample survey methodology one often finds that a frame known to cover approximately all units in the population is one in which sampling is costly while other frames (e.g. special lists of units) are available for cheaper sampling methods. However, the latter usually only cover an unknown or only approximately known fraction of the population. The paper develops a general methodology of utilizing any number of such frames without requiring any prior knowledge as to the extent of their mutual overlap.

2. Some history. (For more history see Sect. 9.)

The technique of multiple frame surveys has been used in the past occasionally and under special circumstances. For example, the 1960 Survey of Agriculture of the Bureau of the Census uses two frames, namely (A) a frame based on the conventional 'area sampling' approach; (B) a frame of farms conceptually and operationally 'associated' with the A-1 listings of the last (1959) Census of Agriculture.* Earlier, the Statistical Laboratory at Iowa State University had used a two frame approach in a small study of 'Effects of Industrialization on Farming' which was carried out for the Department of Economics and Sociology of Iowa State University. Here the two frames consisted of

(A) The customary rural area frame for sampling farm operators.
(B) Employees of the Clinton Motor Company who are also farm operators.

The combined use of these frames proved a successful combination for simulating screening and providing coverage.* There are, no doubt, other occasions when several frames have been used in the past. Although such isolated instances of the use of the method provides valuable experience, it is felt that no systematic methodology for the analysis of such surveys is available.

3. Definitions for two frames.

To fix the ideas, consider first two frames A and B and assume that a sample has been drawn from each frame. The sample designs may be entirely different in the two frames but the following assumptions are made:

(i) Every unit in the population of interest belongs to at least one of the frames.
(ii) It is possible to record for each sampled unit whether or not it belongs to the other frame.

This means we can divide the units of the sample into three ($2^2 - 1$) domains.

Domain (a) The unit belongs to Frame A only
Domain (b) The unit belongs to Frame B only
Domain (ab) The unit belongs to both frames

The units in the population are also conceptually divided into the above domains. We now distinguish four different situations concerning our state of knowledge of the total number of units in the frames and in the domains and of our ability to allocate prescribed sample sizes to the domains.

Schedule 1. Principle situations in multiple frame surveys*

Case No.	Knowledge of population numbers in domains and frames
1	All domain sizes N_a, N_b, N_{ab}, etc. are known
2	All domain sizes N_a, N_b, N_{ab}, etc. are known
3	Domain sizes are not known, but frame sizes are known
4	Neither domain sizes nor frame sizes are known, but the relative magnitude of the frames is known

Case No.	Possibility of fixed sample allocations to domains and frames	Nature of domains
1	It is feasible to allocate prescribed sample sizes to domains	Domains $\equiv$ strata
2	Prescribed sample sizes can only be allocated to frames	Domain $\equiv$ post-strata
3	Prescribed sample sizes can only be allocated to frames	Domains $\equiv$ domains proper
4	Prescribed sample sizes can only be allocated to frames	Domains $\equiv$ domains in populations of unknown size

*For explanation of symbols see Schedule 2.

We now introduce a convenient notation for the two-frame surveys.

Schedule 2. Notation for two frame designs and estimates

	Frame		Domain		
	A	B	a	b	ab
Population number	N_A	N_B	N_a	N_b	N_{ab}
Sample number*	n_A	n_B	n_a	n_b	n'_{ab}, n''_{ab}
Population total	Y_A	Y_B	Y_a	Y_b	Y_{ab}
Population mean	$\bar{Y}_A$	$\bar{Y}_B$	$\bar{Y}_a$	$\bar{Y}_b$	$\bar{Y}_{ab}$
Sample total*	y_A	y_B	y_a	y_b	y'_{ab}, y''_{ab}
Sample mean*	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_a$	$\bar{y}_b$	$\bar{y}'_{ab}$, $\bar{y}''_{ab}$
Cost of sampling unit*	c_A	c_B			

*Applies to case of drawing random samples from both frames.

Note that n'_{ab} and n''_{ab} denote respectively the sub-

*Information from unpublished memoranda.

samples of n_A and n_B respectively which fall into the overlap domain ab. The corresponding means $\bar{y}'_{ab}$ and $\bar{y}''_{ab}$ can only be computed if $n'_{ab} > 0$ and $n''_{ab} > 0$; the same applies to n_a, $\bar{y}_a$ and n_b, $\bar{y}_b$.

4. Formulas for estimation of population totals and means.

In case 1 the estimation problem is reduced to the standard methodology for stratified sampling, whilst in case 4 it will only be possible to estimate population means and not totals. We confine ourselves here to cases 2 and 3. Two approaches leading to identical formulas are possible, viz. (a) the theory of domain estimation, or (b) a special method of weight variables. We use here (b) and introduce the following attributes to units in the two frames:

$$\text{Frame (A)} \quad u_i = \begin{cases} y_i & \text{if } i^{th} \text{ unit is in domain (a)} \\ py_i & \text{if } i^{th} \text{ unit is in domain (ab)} \end{cases}$$

$$\text{Frame (B)} \quad u_i = \begin{cases} y_i & \text{if } i^{th} \text{ unit is in domain (b)} \\ qy_i & \text{if } i^{th} \text{ unit is in domain (ab)} \end{cases}$$

Here p and q are two fixed numbers (to be optimally determined as shown below) with $p+q=1$. We therefore have converted the two frames into two mutually exclusive strata of sizes N_A and N_B by duplicating the N_{ab} units in domain (ab). In stratum (A) there will be N_{ab} units carrying a characteristic $u_i = py_i$ and in stratum (B) there will be N_{ab} units (the 'aliases' of those in stratum (A)) carrying the characteristic $u_i=qy_i$. Clearly Y, the total of the y_i for the original population of $N=N_a+N_{ab}+N_b$ units, is equal to the total U of the u_i for the new population of $N^*=N_a+2N_{ab}+N_b$ units since

$$(1) \quad Y=Y_a+Y_{ab}+Y_b=Y_a+pY_{ab}+qY_{ab}+Y_b=U$$

The standard methodology applicable to the survey designs in Frames (A) and (B) are therefore applicable to obtain estimates of the two stratum-totals (frame-totals) for the variate u_i, their variances and variance estimates. Adding the two will yield the corresponding formulas for the estimation of U and hence Y. To obtain estimates of the population mean $\bar{Y}=Y/N$ apply these formulas to the count variable $x_i=1$ to estimate its total N in analogy to (1) and use the ratio estimate to estimate $Y/N=\bar{Y}$. In case 2 use the device of post stratifying into post strata (a) and (ab) and (b) and (ab) respectively.

5. Formulas for random sampling in both frames in case 2.

We confine ourselves to the simplest case of random sampling in both frames and ignore finite population corrections. In terms of the notation of Schedule 2, using the u-variates in section 4, the (post stratified) estimator of U=Y is given by

$$(2) \quad \hat{Y} = N_a\bar{y}_a+N_{ab}(p\bar{y}'_{ab}+q\bar{y}''_{ab})+N_b\bar{y}_B$$

where the means $\bar{y}_a\bar{y}'_{ab}$ are replaced by $\bar{y}_A$ if either $n_a=0$ or $n'_{ab}=0$, and where likewise the means $\bar{y}_b$ and $\bar{y}_{ab}$ are replaced by $\bar{y}_B$ if either $n_b=0$ or $n''_{ab}=0$. As is well known, under certain restrictions the post stratified estimator Y has a variance approximately equal to that in proportional allocation stratified sampling so that

$$(3) \quad \text{Var } \hat{Y} \doteq \frac{N_A^2}{n_A}\left\{\sigma_a^2\,(1-\alpha) + p^2\sigma_{ab}^2\,\alpha\right\} + \frac{N_B^2}{n_B}\left\{\sigma_b^2\,(1-\beta) + q^2\sigma_{ab}^2\,\beta\right\}$$

where

$$(4) \quad \alpha = N_{ab}/N_A, \; \beta = N_{ab}/N_B,$$

finite population corrections have been ignored and $\sigma_a^2\ \sigma_{ab}^2\sigma_b^2$ are the 'within domain' population variances. Assuming a linear cost function

$$(5) \quad C = c_An_A + c_Bn_B$$

the problem of minimizing (3) as a function of p, n_A and n_B subject to (5) leads to the following specification of the 'optimum' two frame sampling plan:

The optimum value of p is given by a solution of the bi-quadratic

$$(6) \quad c_Ap^2/c_Bq^2 = \frac{\sigma_a^2(1-\alpha) + \alpha\,p^2\sigma_{ab}^2}{\sigma_b^2(1-\beta) + \beta\,q^2\sigma_{ab}^2}$$

With the help of p the optimum sampling fractions are given by

$$(7) \quad \begin{aligned} n_A/N_A &= c\left\{(\sigma_a^2(1-\alpha)+\alpha\,p^2\sigma_{ab}^2)/c_A\right\}^{\frac{1}{2}} \\ n_B/N_B &= c\left\{(\sigma_b^2(1-\beta)+\beta\,q^2\sigma_{ab}^2)/c_B\right\}^{\frac{1}{2}} \end{aligned}$$

with c determined to meet the budget (5). In case of multiple roots of (6) substitution of the alternatives in (3) will select the absolute minimum, except in the rare cases in which the minimum is attained on the boundary of the p, n_An_B space, an occurrence discussed in the literature on optimum allocation in stratified sampling.

Considerable simplifications arise in the important special case in which the A-frame has 100% coverage so that

$$(8) \quad N_{ab} = N_B, \; \beta = 1, \; \sigma_B^2 = \sigma_{ab}^2$$

If these are substituted in (6) the bi-quadratic reduces to the simple equation

$$(9) \quad p^2 = \phi(1-\alpha)\;/\;(\rho-\alpha)$$

where

$$(10) \quad \phi = \sigma_a^2/\sigma_{ab}^2; \; \rho = c_A/c_B; \; \alpha = N_B/N_A$$

In this special case it is possible to compare the variance $\text{Var}(\hat{Y})$ for the optimum design with that of

$$(11) \quad Y' = N_a\bar{y}_a + N_{ab}\bar{y}'_{ab}$$

which is the (post stratified) estimator computed from a simple random sample of size $n_A^* = C/c_A$ drawn from frame A only and requiring an identical budget C. We find for the reduction in variance

$$(12)\quad \mathrm{Var}(\hat{Y})/\mathrm{Var}(Y') = (1 = \frac{q\alpha}{p\rho})^2/(1 + \frac{\alpha q(1+p)}{\rho p^2})$$

with p^2 given by (9) and ϕ, ρ, α given by (10). The reduction in variance for constant cost C (which is of course also the reduction in cost for constant variance) is tabulated in Tables 1, 2 and 3 for the parameter combinations

$$\phi^{-1} = \sigma_B^2/\sigma_a^2 = 1,\ 4,\ 16$$

$$(13)\qquad \rho^{-1} = c_B/c_A = .01,\ .05,\ .10\ (.1)\ .5, 1$$

$$\alpha = N_B/N_A = .5\ (.1)\ .9,\ .95,\ 1$$

The cost reduction may be considerable. Thus for a characteristic situation $\phi^{-1}=16$, $\rho^{-1}=0.2$, $\alpha=0.9$ the reduction is 0.248, i.e. the survey only costs $\frac{1}{4}$ of the 'A-frame only' survey of the same precision. It should, however, be pointed out that the cost (5) only represents the cost depending on the size of the sample and that the (omitted) overhead cost may be larger for the two frame survey because of its more sophisticated design. It should be noted that when $\sigma_B^2/\sigma_a^2 > 1$ that there are two causes for the cost reduction with the two frame design: The first (which is particularly operative when $\alpha = N_B/N_A$ is near 1) is the lower cost of sampling units in B, the second cause (which is more operative when α is near $\frac{1}{2}$) arises when σ_B^2/σ_a^2 is large as the two frame design gets closer to optimum allocation of expected pooled sample sizes to the domains a and ab = B. This situation may arise when the 'cheap' frame B contains the 'important' units, i.e. units with larger and hence more variable y-values. Here the two frame design operates as a 'screening device'. Note in particular in Table 1, that for fixed ρ^{-1} the variance ratio rises from ρ^{-1} to 1 as α is reduced from 1 to 0. Although only the range $0.5 \leq \alpha \leq 1$ is shown it is clear that the function is not monotonically increasing which is a reflection of the above two causes for variance reduction.

6. An alternative cost function in case 2.

The assumptions involved in the cost function (5) imply that the full cost of c_A/unit is incurred for all n_A units sampled from frame A, and likewise for frame B. It may be argued that it may be possible to ascertain at a lower cost c_A' whether or not the unit belongs to domain (a) or (ab), and not to complete the questionnaires for the n_{ab}' units falling into domain ab = B. Since y_{ab}' would then not be available it will be necessary to put p = 0 and q = 1 in (2) and (3) and to modify (5) by replacing the actual cost C by its expectation E(C) and c_A by

$$(14)\qquad c_A^* = c_A(1-\alpha) + c_A'\,\alpha$$

With p = 0 and q = 1 fixed the variance Var $\hat{Y}$ given by (3) must now be minimized as a function of n_A and n_B only subject to a given expected cost E(C). The mathematical formulas are now completely analogous to optimization for strata allocation. We find for the optimum design

$$(15)\quad \frac{n_A}{n_B} = \frac{\sigma_a}{\sigma_B}\sqrt{1-\alpha}\sqrt{\frac{e_B}{e_A}}\frac{N_A}{N_B} = \sqrt{\frac{\phi(1-\alpha)}{\rho}}\ \frac{1}{\alpha}$$

and for the variance reduction

$$(16)\quad \mathrm{Var}\ \hat{Y}/\mathrm{Var}\ Y' = \frac{(1-\alpha\omega)(1+\frac{\alpha}{\sqrt{(1-\alpha)\ \phi\rho}})^2}{(1+\frac{\alpha}{(1-\alpha)\ \phi})}$$

where now

$$(17)\qquad \rho = c_A^*/c_B \text{ and } \omega = 1-(c_A'/c_A)$$

The above reduction may be compared with (12). For ω close to 1 (i.e. c_A^* near to zero) (16) may give a smaller value (larger reduction in variance) than (12). However, in many situations (as for example, with the Census example mentioned in 2.) it is not possible to determine in the field to which domain a sampled unit belongs and with a completion of the questionnaire $c_A' = c_A$, $\omega = 0$, the choice of p = 0 is not optimum and the allocations (7) are preferable.

7. Summary of formulas in case 3.

If domain sizes N_a N_b N_{ab} are not known the ordinary stratified sampling formulas applied to the u-variables in the two strata (frames) of size N_A and N_B must be applied. Thus our estimator of $U = \dot{Y}$ is given by

$$(20)\qquad \dot{Y} = \frac{N_A}{n_A}\left\{y_a + py_{ab}'\right\} + \frac{N_B}{n_B}\left\{y_b + qy_{ab}''\right\}$$

and its variance by

$$(21)\quad \mathrm{Var}(\dot{Y}) = \frac{N_A^2}{n_A}\left\{(1-\alpha)\sigma_a^2 + \alpha\, p^2\sigma_{ab}^2 + \alpha(1-\alpha)(\bar{Y}_a - p\bar{Y}_{ab})^2\right\} + \frac{N_B^2}{n_B}\left\{(1-\beta)\sigma_\beta^2 + \beta\, q^2\sigma_{ab}^2 + \beta(1-\beta)(\bar{Y}_a - q\bar{Y}_{ab})^2\right\}$$

The problem of minimizing $\mathrm{Var}(\dot{Y})$ as a function of p, n_A, n_B subject to a given cost (5) can again be solved and in the present case leads to the optimum allocation formulas

$$(22)\quad n_A^2/N_A^2 = c\left\{\sigma_a^2(1-\alpha) + \sigma_{ab}^2\, p^2\alpha + \alpha(1-\alpha)(\bar{Y}_a - p\bar{Y}_{ab})^2\right\}/c_A$$

$$n_B^2/N_B^2 = c\left\{\sigma_b^2(1-\beta) + \sigma_{ab}^2\, q^2\beta + \beta(1-\beta)(\bar{Y}_b - q\bar{Y}_{ab})^2\right\}/c_B$$

with the constant c to be determined from (5). The allocation formulas (22) involve the value of the optimum p (and q = 1 - p) which must again be determined as the root of a bi-quadratic similar to (6) but not given here. A comparison with an A-only sample design is not appropriate here since the assumption of a frame A with 100% coverage automatically leads to a complete knowledge of the domain sizes from $N_{ab} = N_B$, $N_b = 0$, $N_a = N_A - N_B$ so that the post stratified estimator Y of section 2 should be used.

8. Planned applications.

The small survey on 'Effect of Industrialization on Farming' mentioned in 2 was not designed in accordance with the optimum formulas of section 5 above, in fact it provides an example of the cost situation discussed in 6. Whilst there was not too much difference in the cost values c_A and c_B there was a considerable difference in σ_B^2 and σ_a^2 for most characteristics, in fact it was decided to sample frame B 100 per cent. It is hoped to incorporate illustrative data from this survey and others using the designs here given at the time of publication.

Table 1. Variance reduction in two frame sampling when $\sigma_B^2/\sigma_a^2 = 16$

Sampling cost ratio	N_B/N = proportion of population in cheap frame						
c_B/c_A	.5	.6	.7	.8	.9	.95	1
.01	.096	.076	.059	.045	.031	.024	.010
.05	.154	.134	.118	.102	.086	.075	.050
.10	.206	.188	.174	.160	.143	.131	.100
.20	.288	.278	.269	.261	.248	.237	.200
.30	.359	.356	.355	.353	.347	.338	.300
.40	.423	.428	.435	.440	.441	.436	.400
.50	.483	.496	.510	.524	.533	.532	.500
1.00	.735	.784	.836	.889	.944	.972	1

Table 2. Variance reduction in two frame sampling when $\sigma_B^2/\sigma_a^2 = 4$

Sampling cost ratio	N_B/N = proportion of population in cheap frame						
c_B/c_A	.5	.6	.7	.8	.9	.95	1
.01	.259	.201	.152	.108	.066	.044	.010
.05	.340	.284	.234	.186	.137	.107	.050
.10	.404	.352	.304	.257	.205	.172	.100
.20	.500	.456	.415	.372	.322	.287	.200
.30	.576	.540	.507	.472	.426	.393	.300
.40	.640	.613	.588	.561	.523	.493	.400
.50	.696	.678	.661	.642	.614	.589	.500
1.00	.900	.914	.932	.953	.976	.988	1

Table 3. Variance reduction in two frame sampling when $\sigma_B^2/\sigma_a^2 = 1$

Sampling cost ratio	N_B/N = proportion of population in cheap frame						
c_B/c_A	.5	.6	.7	.8	.9	.95	1
.01	.571	.477	.379	.276	.164	.101	.010
.05	.656	.573	.482	.381	.260	.186	.050
.10	.718	.645	.562	.465	.344	.263	.100
.20	.800	.742	.674	.589	.475	.392	.200
.30	.857	.812	.757	.686	.582	.503	.300
.40	.900	.866	.824	.765	.676	.604	.400
.50	.933	.909	.877	.832	.759	.695	.500
1.00	1	1	1	1	1	1	1

9. The Bureau of the Census Survey of Retail Stores.

It has been pointed out to us that mention should be made of the 'Sample Survey of Retail Stores' by the Bureau of the Census (1949) which is perhaps one of the largest and earliest instances of the combined use of a list-frame and an area frame. For a description of this survey see Hansen, Hurwitz and Madow (1953) 'Sample Survey Methods and Theory', Vol. 1 (pp. 515-558). This survey carefully avoids the sampling of list-units encountered in the sampled area segments and therefore follows essentially the method described in 6., although the area sample design is multi-stage. In the description of this survey it is not discussed whether the cost c_a' of screening out list units from the area segment warrants their omission and the use of weight coefficients p=0 and q=1 in place of the optimum p and q of Section 5. In fact the method of optimum weight coefficients p and q has to the best of our knowledge never been used. Cost c_a' consists of recognizing a place of business located within segment boundaries making sure that it is identical with an establishment whose address is mentioned on the list and then discarding it from further interview.

An example of a similar type is discussed on pp. 327-8 of the above book, where a table of variance reduction is given for a special situation which corresponds closely to our case 1 in Schedule 2, which follows standard stratification.