THE ESTIMATION OF DOMAIN SIZES WHEN SAMPLING FRAMES ARE INTERLOCKING

Robert S. Cochran, Statistics Department, University of Wyoming

I. INTRODUCTION

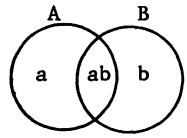
The concept of interlocking sampling frames has been discussed by Hartley (1962) and by Cochran (1964). In these papers estimates of general y characteristics have been presented for situations where two frames overlap and the population frequency in each of the three domains is known. Hartley (1962) also discussed the estimation of the population total for y when the domain frequencies are not known. The purpose of this paper is to consider the estimation of the number in the domains created by the overlapping of two or three frames.

Bryant and King (1960) treated the problem when three frames overlap by using the modified minimum chi-square estimation technique. In the research for this paper some corrections in their approach were made and some comparisons were also made between their estimators and those obtained using multiple-frame procedures.

II. SIMPLE RANDOM SAMPLING FROM EACH OF TWO FRAMES

A. Sample Size Given and Weights to be Determined

The two frames being sampled are frames A and B with frequencies N_A and N_B respectively. From these frames two independent random samples of size n_A and n_B are selected without replacement. Because there are some members of the population in both frames the three domains a, b and ab are created by the use of the two frames. Those members of the population that are just in frame A are in domain a, those that are just in frame B are in domain b, and those that belong to two frames are in domain ab.



After the samples have been drawn the elements selected are classified into their proper domain. In this way we have

$$n_a + n'_{ab} = n_A \quad \text{and} \quad n''_{ab} + n_b = n_B .$$

From this information the estimates of N_a, N_b and N_{ab} are to be computed. It is obvious that since N_A and N_B are known it is only necessary to concentrate on the estimation of one of these with the other two being obtained by subtraction.

Without loss of generality we will estimate N_{ab} directly and obtain the estimates of N_a and N_b by subtraction as

$$\hat{N}_a = N_A - \hat{N}_{ab} \quad \text{and} \quad \hat{N}_b = N_B - \hat{N}_{ab} .$$

The number of distinct elements in the population can also be estimated as

$$\hat{N} = N_A + N_B - \hat{N}_{ab} .$$

When confronted with the information from two independent samples about the relative frequency of the overlap area it seems only natural that the best way to estimate the frequency of this area is by combining this information. From frame A there is

$$\hat{N}'_{ab} = \frac{N_A}{n_A} n'_{ab}$$

and from frame B there is

$$\hat{N}''_{ab} = \frac{N_B}{n_B} n''_{ab} .$$

If the sampling fractions, n_i/N_i, are small enough and the populations are large enough these estimates have variances

$$V(\hat{N}'_{ab}) = \frac{N_A^2}{n_A}\left(\frac{N_{ab}}{N_A}\right)\left(1 - \frac{N_{ab}}{N_B}\right) = \frac{N_A^2}{n_A}\alpha(1-\alpha)$$

and

$$V(\hat{N}''_{ab}) = \frac{N_B^2}{n_B}\left(\frac{N_{ab}}{N_B}\right)\left(1 - \frac{N_{ab}}{N_B}\right) = \frac{N_B^2}{n_B}\beta(1-\beta) .$$

Combining these two independent estimators yields the multiple frame estimator

$$\hat{N}_{ab} = p\hat{N}'_{ab} + q\hat{N}''_{ab}, \quad \text{where} \quad p+q = 1 .$$

While it has been developed here as a weighted average of two independent estimators it can also be developed by using the multiple frame approach of Hartley (1962).

Using the well-known principle of linear functions of independent random variables the variance of $\hat{N}_{ab}$ is easily seen to be

$$V(\hat{N}_{ab}) = p^2V(\hat{N}'_{ab}) + q^2V(\hat{N}''_{ab}) = p^2\frac{N_A^2}{n_A}\alpha(1-\alpha) + q^2\frac{N_B^2}{n_B}\beta(1-\beta) .$$

The value of p that will minimize this variance is

$$p_0 = \frac{V(\hat{N}''_{ab})}{V(\hat{N}'_{ab}) + V(\hat{N}''_{ab})} .$$

Cochran (1965) shows some results of using a nonoptimum p in terms of the loss in precision for the estimator.

Bryant and King (1960) did not deal with two-frame situations. However, Cochran (1965) did use their procedure and came up with the result that for two frames the estimators are algebraically equivalent and have the same estimates of their variances when p_0 was estimated from the sample information.

B. Sample Size and Weights to be Determined

In the application of these techniques to most surveys the estimation of the number in the separate domains is only one of several pieces of information desired from the survey. With this in mind the sample sizes drawn are usually selected to give maximum information on some other variable or to satisfy some other restraint. However, when the estimation of the frequency in each of the several domains is either the only quantity of interest or is the most important quantity some interesting results can be given.

In the case of two frames where the quantity N_{ab} is to be estimated it was previously given that

$$\hat{N}_{ab} = p\hat{N}'_{ab} + q\hat{N}''_{ab}$$

and

$$V(\hat{N}_{ab}) = p^2 \frac{N_A^2}{n_A} \alpha(1-\alpha) + q^2 \frac{N_B^2}{n_B} \beta(1-\beta) .$$

Assuming n_A and n_B to have been previously determined the optimum value for p becomes

$$p_0 = \frac{\frac{N_B^2}{n_B} \beta(1-\beta)}{\frac{N_A^2}{n_A} \alpha(1-\alpha) + \frac{N_B^2}{n_B} \beta(1-\beta)} .$$

Now setting the partial derivative of $V(N_{ab})$ with respect to n_A and n_B subject to a cost condition

$$C = n_A C_A + n_B C_B$$

equal to zero yields

$$n_A = p\left(\frac{N_A^2 \alpha(1-\alpha)}{C_A}\right)^{1/2} \text{ and } n_B = q\left(\frac{N_B^2 \beta(1-\beta)}{C_B}\right)^{1/2} .$$

The constant λ can be shown to be

$$\lambda = C^{-1}[p(N_{ab}N_a C_A)^{1/2} + q(N_{ab}N_b)^{1/2}]$$

Substituting the optimum n_A and n_B into the expression for the optimum p yields

$$p = \frac{p[\frac{1}{N_A^2 \alpha(1-\alpha)C_A}]^{1/2}}{p[\frac{1}{N_A^2 \alpha(1-\alpha)C_A}]^{1/2} + q[\frac{1}{N_B^2 \beta(1-\beta)C_B}]^{1/2}} .$$

This expression when solved for p has two solutions, p = 0 and p = 1, unless

$$N_B^2 \beta(1-\beta)C_B = N_A^2 \alpha(1-\alpha)C_A .$$

In this unlikely case any value for p, $0 \leq p \leq 1$ will be a solution.

In order to determine conditions that will indicate which of the values, p = 0 or p = 1, actually gives the minimum variance consider the variance equation under each of these choices.

When p = 1 the sample size should be

$$n_A = \frac{C}{C_A}$$

and the variance becomes

$$V_1 = \frac{N_A^2 \alpha(1-\alpha)C_A}{C}$$

When p = 0 the sample size should be

$$n_B = \frac{C}{C_B}$$

and the variance becomes

$$V_0 = \frac{N_B^2 \beta(1-\beta)C_B}{C} .$$

Thus the question can be settled by considering the relationship between $\frac{1-\beta}{\beta} C_B$ and $\frac{1-\alpha}{\alpha} C_A$ or $N_b C_B$ and $N_a C_A$. When $\frac{1-\beta}{\beta} C_B > \frac{1-\alpha}{\alpha} C_A$, then V_1 is the smaller. When all sampling costs are the same ($C_A = C_B$) the above can be shown to imply that $N_B > N_A$. Therefore with equal costs of sampling, sample entirely from the smallest frame.

When the sampling costs are not the same the relationship between $C_B N_b$ and $C_A N_a$ can sometimes be derived from the relationships between C_B and C_A, N_B and N_A, and $C_B N_B$ and $C_A N_A$. Whenever

$$N_A > N_B \quad \text{and} \quad C_A \geq C_B$$

then

$$N_A C_A > N_B C_B \quad \text{and} \quad N_a C_A > N_b C_B$$

and the obvious decision is to sample from the small cheap frame. Whenever

$$C_B > C_A \quad \text{and} \quad N_A C_A \geq N_B C_B \quad \text{then} \quad C_A N_a > C_B N_b .$$

The third possibility is

$$C_B > C_A \quad \text{and} \quad N_B C_B > N_A C_A .$$

In this case the result depends upon the unknown N_{ab}. $C_A N_a$ will be larger whenever

$$\frac{N_B C_B - N_A C_A}{C_B - C_A} < N_{ab} .$$

III. SIMPLE RANDOM SAMPLING FROM THREE FRAMES

When sampling from three frames, A, B and C, there are seven domains. Of these, it is only necessary to directly estimate the number of units in four. The number of units in the remaining three can be estimated by subtraction from the known domain sizes. Without loss of generality let these four be N_{ab}, N_{ac}, N_{bc} and N_{abc}, the number of units in the areas of overlap.

Using an obvious extension of the notation and procedures of the two-frame case above the following estimates are obtained:

$$\hat{N}_{ab} = p_{ab} \frac{N_A}{n_A} n'_{ab} + q_{ab} \frac{N_B}{n_B} n''_{ab}$$

$$\hat{N}_{ac} = p_{ac} \frac{N_A}{n_A} n'_{ac} + q_{ac} \frac{N_C}{n_C} n''_{ac}$$

$$\hat{N}_{bc} = p_{bc} \frac{N_B}{n_B} n'_{bc} + q_{bc} \frac{N_C}{n_C} n''_{bc}$$

and

$$\hat{N}_{abc} = p_A \frac{N_A}{n_A} n'_{abc} + p_B \frac{N_B}{n_B} n''_{abc} + p_C \frac{N_C}{n_C} n'''_{abc} .$$

The variances of the quantities are

$$V(\hat{N}_{ab}) = p_{ab}^2 \frac{N_A^2}{n_A} \alpha_1(1-\alpha_1) + q_{ab}^2 \frac{N_B^2}{n_B} \alpha_2(1-\alpha_2),$$

$$\alpha_1 = \frac{N_{ab}}{N_A}; \quad \alpha_2 = \frac{N_{ab}}{N_B}$$

$$V(\hat{N}_{ac}) = p_{ac}^2 \frac{N_A^2}{n_A} \gamma_1(1-\gamma_1) + q_{ac}^2 \frac{N_C^2}{n_C} \gamma_2(1-\gamma_2),$$

$$\gamma_1 = \frac{N_{ac}}{N_A}; \quad \gamma_2 = \frac{N_{ac}}{N_C}$$

$$V(\hat{N}_{bc}) = p_{bc}^2 \frac{N_B^2}{n_B} \beta_1(1-\beta_1) + q_{bc}^2 \frac{N_C^2}{n_C} \beta_2(1-\beta_2),$$

$$\beta_1 = \frac{N_{bc}}{N_B}; \quad \beta_2 = \frac{N_{bc}}{N_C}$$

and

$$V(\hat{N}_{abc}) = p_A^2 \frac{N_A^2}{n_A} \delta_1(1-\delta_1) + p_B^2 \frac{N_B^2}{n_B} \delta_2(1-\delta_2) + p_C^2 \frac{N_C^2}{n_C} \delta_3(1-\delta_3), \qquad \delta_1 = \frac{N_{abc}}{N_A};$$

$$\delta_2 = \frac{N_{abc}}{N_B}; \quad \delta_3 = \frac{N_{abc}}{N_C} .$$

The values of the p's that minimize these variances are:

$$p_{ab} = \frac{V(\hat{N}''_{ab})}{V(\hat{N}'_{ab}) + V(\hat{N}''_{ab})}, \qquad q_{ab} = 1-p_{ab}$$

$$p_{ac} = \frac{V(\hat{N}''_{ac})}{V(\hat{N}_{ac}) + V(\hat{N}''_{ac})}, \qquad q_{ac} = 1-p_{ac}$$

$$p_{bc} = \frac{V(\hat{N}''_{bc})}{V(\hat{N}'_{bc}) + V(\hat{N}''_{bc})}, \qquad q_{bc} = 1-p_{bc}$$

$$p_A = \frac{\frac{1}{V(\hat{N}'_{abc})}}{\frac{1}{V(\hat{N}'_{abc})} + \frac{1}{V(\hat{N}''_{abc})} + \frac{1}{V(\hat{N}'''_{abc})}} = \frac{V(\hat{N}''_{abc})\, V(\hat{N}'''_{abc})}{V(\hat{N}'_{abc})V(\hat{N}''_{abc}) + V(\hat{N}'_{abc})V(\hat{N}'''_{abc}) + V(\hat{N}''_{abc})V(\hat{N}'''_{abc})}$$

$$p_B = \frac{\frac{1}{V(\hat{N}''_{ab})}}{\frac{1}{V(\hat{N}'_{abc})} + \frac{1}{V(\hat{N}''_{abc})} + \frac{1}{V(\hat{N}'''_{abc})}}$$

$$p_C = \frac{\frac{1}{V(\hat{N}'''_{abc})}}{\frac{1}{V(\hat{N}'_{abc})} + \frac{1}{V(\hat{N}''_{abc})} + \frac{1}{V(\hat{N}'''_{abc})}} .$$

The estimation of quantities such as N_a can now be carried out by subtraction of the estimates of N_{ab}, N_{ac}, and N_{abc} from the known frame size N_A,

$$\hat{N}_a = N_A - (\hat{N}_{ab} + \hat{N}_{ac} + \hat{N}_{abc}) .$$

Cochran (1965) shows the variance of $\hat{N}_a$ to be

$$\begin{aligned} V(\hat{N}_a) = \frac{N_A^2}{n_A} \Big\{ & p_{ab}^2 \alpha_1(1-\alpha_1) + p_{ac}^2 \gamma_1(1-\gamma_1) \\ & + p_A^2 \delta_1(1-\delta_1) - 2\, 2p_{ab}\, p_{ac}\, \alpha_1\gamma_1 \\ & - 2\, p_{ab}p_A\alpha_1\delta_1 - 2\, p_{ac}p_A\gamma_1\delta_1 \Big\} \\ & + \frac{N_B^2}{n_B} \Big\{ q_{ab}^2 \alpha_2(1-\alpha_2) + p_B^2 \delta_2(1-\delta_2) \\ & - 2\, p_B\, q_{ab}\, \alpha_2\, \delta_2 \Big\} \\ & + \frac{N_C^2}{n_C} \Big\{ q_{ac}^2 \gamma_2(1-\gamma_2) + p_C^2 \delta_3(1-\delta_3) \\ & - 2\, p_C\, q_{ac}\, \gamma_2\, \delta_3) \Big\} . \end{aligned}$$

As an example of the multiple-frame approach consider the 1964 data of the following table.

License Frame	1964 Population			Est.	Var.
	Deer	Elk	A'lope		
Deer only	1,637			22,425	
Elk only		765		7,277	
Antelope only			278	1,915	
Deer-Elk	1,023	1,402		13,222	44,830
Deer-Antelope	353		549	4,454	16,704
Elk-Antelope		107	115	954	4,171
Deer-Elk-Ant.	48	720	768	6,372	14,004
Total Sample	3,497	2,994	1,710		
Pop. Size	46,473	27,825	13,695	56,619	

For the deer-elk overlap

$$\hat{N}'_{de} = \frac{46,473}{3,497} \cdot 1023 = 13,596$$

$$\hat{N}''_{de} = \frac{27,825}{2,994} \cdot 1402 = 13,030$$

$$v(\hat{N}'_{de}) = 145,074; \quad v(\hat{N}''_{de}) = 64,416$$

$$p_{de} = .34 .$$

For the deer-elk-antelope overlap

$$\hat{N}'_{dea} = \frac{46,473}{3,497} \cdot 484 = 6,433$$

$$\hat{N}''_{dea} = \frac{27,825}{2,994} \cdot 720 = 6,692$$

$$\hat{N}'''_{dea} = \frac{13,695}{1,710} \cdot 768 = 6,151$$

$$v(\hat{N}'_{dea}) = 74,359; \quad v(\hat{N}''_{dea}) = 47,168;$$

$$v(\hat{N}'''_{dea}) = 27,146; \; p_D = .9; \; p_E = .31; \; p_A = .50 .$$

The approach of Bryant and King (1960) leads to modified minimum chi-square estimates for the three-frame case, the solution of four simultaneous equations in four unknowns. In matrix notation this is

$$A\hat{N} = Y$$

$$\hat{N} = A^{-1} Y$$

The A matrix of coefficients is made up of the partial derivations of

$$\chi^2 = \frac{(n'_{ab} - \frac{n_A}{N_A} N_{ab})^2}{n'_{ab}} + \frac{(n''_{ab} - \frac{n_B}{N_B} N_{ab})^2}{n''_{ab}}$$
$$+ \frac{(n'_{ac} - \frac{n_A}{N_A} N_{ac})^2}{n'_{ac}} + \frac{(n''_{ac} - \frac{n_C}{N_C} N_{ac})^2}{n''_{ac}}$$
$$+ \frac{(n'_{bc} - \frac{n_B}{N_B} N_{bc})^2}{n'_{bc}} + \frac{(n''_{bc} - \frac{n_C}{N_C} N_{bc})^2}{n''_{bc}}$$
$$+ \frac{(n'_{abc} - \frac{n_A}{N_A} N_{abc})^2}{n'_{abc}} + \frac{(n''_{abc} - \frac{n_B}{N_B} N_{abc})^2}{n''_{abc}}$$
$$+ \frac{(n'''_{abc} - \frac{n_C}{N_C} N_{abc})^2}{n'''_{abc}}$$

with respect to N_{ab}, N_{ac}, N_{bc} and N_{abc}. The Y vector contains the constants arising from the differentiation process. They derive the variance of these estimates to be of the form

$$[F]\ [\sigma_{rs}]\ [F]'$$

where

$$[F] = F'_{n'_{ab}}, F'_{n''_{ab}}, \ldots, F'_{n'''_{abc}}\ {}_{4x9}$$

$$F'_{n_i} = -A^{-1} \frac{\partial A}{\partial n_i} A^{-1} Y + A^{-1} \frac{\partial Y}{\partial n_i}$$

and $[\sigma_{rs}]$ is the 9 x 9 variance-covariance matrix for the number of observations appearing in the overlap areas.

No specific analytical comparisons were made between these estimates and their variances with the multiple-frame type of estimates. However, some numerical comparisons were made using information from the 1960 through 1964 big game studies conducted for the Wyoming Game and Fish Commission by the University of Wyoming.

The figure for these five years indicates close agreement between the two estimators. However, in all but one instance (1962, E-2D) the estimate of the variance of the multiple-frame estimate was less than the estimate of the variance of the minimum chi-square estimate. In some cases there was an appreciable gain using the multiple-frame estimator.

Year	Class	Estimate		Variances	
		Multiple Frame	Minimum χ^2	Multiple Frame	Minimum χ^2
1960	D-E	14,333	14,397	58,452	66,461
	D-2D	1,446	1,408	8,631	12,500
	E-2D	454	439	2,420	2,720
	D-E-2D	1,466	1,442	5,545	8,555
1961	D-E	16,006	15,865	68,937	73,259
	D-2D	1,332	1,417	7,792	10,660
	E-2D	243	237	1,882	2,194
	D-E-2D	1,179	1,172	5,040	7,883
1962	D-E	11,633	11,272	108,013	115,100
	D-2D	1,048	1,035	11,861	12,583
	E-2D	2,124	2,015	20,893	15,544
	D-E-2D	1,475	1,332	11,833	14,017
1963	D-E	12,179	11,852	81,635	136,079
	D-A	5,629	5,794	27,981	125,882
	E-A	875	880	6,316	8,567
	D-E-A	6,104	6,058	22,626	91,786
1964	D-E	13,222	13,450	44,830	86,697
	D-A	4,454	4,379	16,704	61,979
	E-A	954	925	4,171	5,350
	D-E-A	5,372	6,299	14,004	57,822

IV. LITERATURE CITED

Bryant, Edward C. and King, Donald W. (1960). "Estimation from Populations Identified by Overlapping Sample Frame." Unpublished paper presented at the American Statistical Association meeting, Palo Alto, California.

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

_____. (1965). "Theory and Application of Multiple Frame Surveys." Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.

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