ON THE DERIVATION OF OPTIMUM ALLOCATION FORMULAS IN STRATIFIED MULTI-STAGE SAMPLING BY THE USE OF THE CAUCHY INEQUALITY

J. C. Koop, North Carolina State College, Raleigh, North Carolina

INTRODUCTION

Without using the differential calculus, Neyman [1] and Sukhatme [2], in his well-known textbook on statistical sampling, have derived optimum allocation formulas for their sampling problems by purely algebraic arguments. Stuart [3] applied Cauchy's inequality to simplified variance and cost functions to obtain such formulas and illustrated the basic simplicity of the method.

This well-known inequality may be stated as follows. If a_i and b_i $(i = 1,2,....r)$ are real numbers then[1]

$$(a_1^2 + ... + a_r^2)(b_1^2 + ... + b_r^2) \geq (a_1b_1 + ... + a_rb_r)^2 \quad (1)$$

and equality is only attained if

$$a_1/b_1 = a_2/b_2 = ... = a_r/b_r. \quad (2)$$

When equality in (1) is attained, under the conditions stated at (2), then the left hand side of (1) is necessarily a minimum.

In this expository paper the application of (1) to variance functions and certain cost functions appropriate to more involved sample designs in stratified multi-stage sampling, and some related, will be considered. By letting $\sum a^2$ play the role of the variance function, and $\sum b^2$ that of the cost function, it will appear that the set of conditions specified by (2) in respect of the unknown sample sizes for each stratum and for each step in sampling will lead to equations from which these unknowns can be solved for each of the situations when either the over-all cost of survey or the variance (for a character of interest) is fixed. These solutions give the optima.

The presentation of such a paper appears to be warranted as so far the simplicity of this method together with its ease of application for more ramified sample designs is still not so well known.

STRATIFIED SAMPLING IN TWO STAGES

We shall consider two cases here, namely (i) when the first- and second-stage units are sampled with equal probabilities and without replacement at each stage, (ii) when the first-stage units are sampled with unequal probabilities (e.g. probabilities proportional to certain measures of size) and with replacement and the second-stage units with equal probabilities and without replacement. In all problems the estimation of population totals will be considered. This will be sufficient to illustrate the method. For theoretical development of formulas reference is made to textbooks on statistical sampling.

<u>Case (i).</u> There are L strata and it is desired to estimate the total

$$T = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} x_{hij}. \quad (3)$$

where x_{hij} is a variate of interest of the j^{th} second stage unit $(j = 1,2,...N_{hi})$ of the i^{th} first stage unit $(i = 1,2,...N_h)$ of the h^{th} stratum $(h = 1,2,...L)$. The best unbiased linear estimate based on n_h first-stage units $(h = 1,2,...L)$ and n_{hi} second-stage units $(i = 1,2,...n_h)$ is given by

$$T' = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{n_{hi}} x_{hij}. \quad (4)$$

The variance of T' is given by

$$V(T') = \sum_{h=1}^{L} N_h^2 \left[\frac{S_h^2}{n_h}(1 - \frac{n_L}{N_L}) + \frac{1}{N_h n_h} \sum_{i=1}^{N_h} N_{hi}^2 \frac{S_{hi}^2}{n_{hi}} (1 - \frac{n_{hi}}{N_{hi}}) \right] \quad (5)$$

where in the h^{th} stratum S_h^2 is the variance of the totals for the first-stage units and S_{hi}^2 is the variance of the second-stage units in the i^{th} first-stage unit each with divisor N_h-1 and N_{hi}-1 respectively. The expected cost C appropriate to variance function (5) is given by

$$C = C_o + \sum_{h=1} (c_h n_h + \sum_{i=1}^{N_h} \frac{C_{hi}}{N_h} n_h n_{hi})$$

or

$$C - C_o = \sum_h c_h n_h + \sum_h \sum_i \frac{c_{hi}}{N_h} n_h n_{hi} \quad (6)$$

where C_o is the fixed over-all cost, c_h is the cost per first-stage unit, and c_{hi} is the cost per second-stage unit in the i^{th} first-stage unit in stratum h. The above cost function (6) is sufficiently general. Others more involved do not lead to easy solutions.

To apply Cauchy's inequality, (5), after rearrangement of terms, is rewritten as

$$V(T') + \sum_{h=1}^{L} N_h S_h^2 = \sum_{h=1}^{L} \frac{N_h^2 Z_h^2}{n_h} + \sum_{h=1}^{L} \sum_{i=1}^{N_h} \frac{N_h N_{hi}^2 S_{hi}^2}{n_h n_{hi}} , \quad (5.1)$$

where $Z_h^2 = S_h^2 - \sum_{i=1}^{N_h} N_{hi} S_{hi}^2 / N_h$. It will be noted that the left hand sides of both (6) and (5.1) do not involve the unknowns explicitly. Let us assume that all Z_h^2 are positive. Now applying the inequality we have

$$(V(T') + \sum_h N_h S_h^2)(C - C_o)$$

$$\geq \left[\sum_h \frac{N_h Z_h}{\sqrt{n_h}} \cdot \sqrt{c_h n_h} + \sum_h \sum_i \frac{\sqrt{N_h} \cdot N_{hi} S_{hi}}{\sqrt{n_h n_{hi}}} \cdot \sqrt{\frac{c_{hi}}{N_h} n_h n_{hi}} \right]^2 \quad (7)$$

$$\text{i.e.} \geq \left[\sum_h \left(\sqrt{c_h} N_h Z_h + \sum_i \sqrt{c_{hi}} N_{hi} S_{hi} \right) \right]^2. \quad (7.1)$$

It should be noted that

$$N_h Z_h / \sqrt{n_h}, \quad h = 1,2,..L$$

and

$$\sqrt{N_h} N_{hi} S_{hi} / \sqrt{n_h n_{hi}}, \quad h=1,2,..L \text{ and } i=1,2,...N_h,$$

play the role of the a's. Similarly $\sqrt{c_n n_h}$ and $\sqrt{c_{hi} n_h n_{hi} / N_h}$ for relevant h and (h,i) play the role of the b's. On the right hand side of (7), the value of the expression in square brackets does not involve the unknown sample sizes. This value is the minimum value and is attained when

$$\frac{\frac{N_h Z_h}{\sqrt{n_h}}}{\sqrt{c_h n_h}} = \frac{\frac{\sqrt{N_h} N_{hi} S_{hi}}{\sqrt{n_h n_{hi}}}}{\sqrt{\frac{c_h n_h n_{hi}}{N_h}}} = \lambda, \quad (8)$$

where λ is a constant, for all relevant h and (h,i) noted above. There are $L + \sum_{h=1}^{L} N_h$ such equations. From (8) we find the optimum values to be

$$n_{hi} = \sqrt{\frac{c_h}{c_{hi}}} \cdot \frac{N_{hi} S_{hi}}{Z_h}, \quad (9)$$

$$(h = 1,2,..L, \; i=1,2,...N_h)$$

and

$$n_h = \frac{1}{\lambda} \cdot \frac{N_h Z_h}{\sqrt{c_h}} \cdot (h = 1,2,..L). \quad (10)$$

From (9) we see that for a given stratum the n_{hi} are independent of the parameters relating to other strata. By fixing either the variance or the cost, the constant $1/\lambda$ can be determined from (5.1) or (6) as the case may be by substituting for n_{hi} and n_h given by equations (9) and (10). Let $V(T')$ be fixed at V_o, i.e. $V(T') = V_o$. Then from (5.1) it will be found that

$$\frac{1}{\lambda} = \sum_{h=1}^{L} \sqrt{c_h} \left(Z_h + \frac{1}{\sqrt{c_h}} \sum_{i=1}^{N_h} \sqrt{c_{hi}} N_{hi} S_{hi} \right) / \left(V_o + \sum_h N_h S_h^2 \right) \quad (11)$$

and thus

$$n_h = \frac{N_h Z_h}{\sqrt{c_h}} \sum_h \sqrt{c_h} \left(Z_h + \frac{1}{\sqrt{c_h}} \sum_i \sqrt{c_{hi}} N_{hi} S_{hi} \right) / \left(V_o + \sum_h N_h S_h^2 \right). \quad (12)$$

The conclusions which emerge from an inspection of the formulas for n_{hi} and n_h in (9) and (12) respectively (barring the exceptional situations for which Z_h^2 may be negative) are quite familiar and will not be repeated. When the over-all cost $C = C'$, then from (6) we will find

$$\frac{1}{\lambda} = (C' - C_o) / \sum_h \left(\sqrt{c_h} N_h Z_h + \sum_i \sqrt{c_{hi}} N_{hi} S_{hi} \right),$$

and thus

$$n_h = \frac{C' - C_o}{\sqrt{c_h}} \cdot \frac{N_h Z_h}{\sum_h \left(\sqrt{c_h} N_h Z_h + \sum_i \sqrt{c_{hi}} N_{hi} S_{hi} \right)}. \quad (13)$$

Finally, from (7.1) the minimum expected variance or cost is obtained when either cost or variance is fixed respectively.

Case (ii). In this sampling plan the first-stage units are sampled with unequal probabilities and with replacement. Let p_{hi} be the probability of selection of the i^{th} first-stage unit in the h^{th} stratum where $\sum_{i=1}^{N_h} p_{hi} = 1$ for $h = 1,2,....L$. In this case a simple unbiased estimate T'' of T will be given by

$$T'' = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{N_{hi}}{p_{hi} n_{hi}} \sum_{j=1}^{n_{hi}} x_{hij}, \quad (14)$$

with variance

$$V(T'') = \sum_{h=1}^{L} \left[\frac{\sigma_h^2}{n_h} + \frac{1}{n_h} \sum_{i=1}^{N_h} \frac{N_{hi}^2}{p_{hi}} \cdot \frac{S_{hi}^2}{n_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) \right], \quad (15)$$

where

$$\sigma_h^2 = \sum_{i=1}^{N_h} p_{hi} \left(\frac{x_{hi}}{p_{hi}} - x_h \right)^2, \quad (h = 1,2,...L)$$

in which

$$\left. \begin{array}{l} x_{hi} = \sum_{j=1}^{N_{hi}} x_{hij}, \\ \text{and} \\ x_h = \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} x_{hij} \end{array} \right\} (h = 1,2,...L)$$

The remaining notation is the same as for Case (i). The variance function (15) can be rewritten as

$$V(T'') = \sum_{h=1}^{L} \left[\frac{U_h^2}{n_h} + \sum_{i=1}^{N_h} \frac{N_{hi}^2 S_{hi}^2}{p_{hi} n_h n_{hi}} \right], \tag{16}$$

where $U_h^2 = \sigma_h^2 - \sum_{i=1}^{N_h} (N_{hi} S_{hi}^2)/p_{hi}$, to facilitate the application of the inequality. It will be assumed that all U_h^2 are all positive. An expected cost function C_1 appropriate to (16) is

$$C_1 = C_o + \sum_{h=1}^{L} (c_h n_h + n_h \sum_{i=1}^{N_h} p_{hi} c_{hi} n_{hi}). \tag{17}$$

Proceeding on the same lines as in Case (i) the optima n_{hi} are given by

$$n_{hi} = \sqrt{\frac{c_h}{c_{hi}}} \cdot \frac{N_{hi} S_{hi}}{p_{hi} U_i} \cdot (h=1,2,..L; i=1,2,..N_h) \tag{18}$$

For fixed cost C_1' or fixed variance V_o the optima n_h are given respectively by

$$\left. \begin{aligned} n_h &= \frac{C_1' - C_o}{\sqrt{c_h}} \cdot \frac{U_h}{\sum_h (\sqrt{c_h}\, U_h + \sum_i \sqrt{c_{hi}}\; N_{hi} S_{hi})} \\ &\text{and} \\ n_h &= \frac{1}{\sqrt{c_h}} \cdot \frac{U_h}{V_o} \sum_h (\sqrt{c_h}\, U_h + \sum_i \sqrt{c_{hi}}\; N_{hi} S_{hi}) \end{aligned} \right\} \cdot (h = 1,2,.L) \tag{19}$$

SAMPLING IN THREE OR MORE STAGES

The extension of the problem considered under Case (i) to three or more stages will be given. Let x_{hijk} be the variate of interest of the k^{th} third-stage unit ($k = 1,2,...N_{hij}$). The remaining notation is as before. Let n_{hij} third-stage units be sampled without replacement and with equal probabilities in the j^{th} second-stage unit of the i^{th} first-stage unit of the h^{th} stratum. The total T will be given by

$$T = \sum_h \sum_i \sum_j \sum_k x_{hijk}, \tag{20}$$

the best unbiased linear estimate of which will be given by

$$T' = \sum_h \frac{N_h}{n_h} \sum_i \frac{N_{hi}}{n_{hi}} \sum_j \frac{N_{hij}}{n_{hij}} \sum_k x_{hijk} \tag{21}$$

with variance

$$V(T') = \sum_{h=1}^{L} N_h^2 \left[\frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) + \frac{1}{N_h n_h} \sum_{i=1}^{N_h} N_{hi}^2 \cdot \left\{ \frac{S_{hi}^2}{n_{hi}}\left(1 - \frac{n_{hi}}{N_{hi}}\right) + \frac{1}{N_{hi} n_{hi}} \sum_{j=1}^{N_{hi}} N_{hij}^2 \cdot \frac{S_{hij}^2}{n_{hij}}\left(1 - \frac{n_{hij}}{N_{hij}}\right) \right\} \right] \tag{22}$$

The variance function (22) can be rewritten as

$$V(T') + \sum_{h=1}^{L} N_h S_h^2 = \sum_h \frac{N_h^2 Z_h^2}{n_h} + \sum_h \sum_i \frac{N_h N_{hi}^2 Z_{hi}^2}{n_h n_{hi}} + \sum_h \sum_i \sum_j \frac{N_h N_{hi} N_{hij}^2 S_{hij}^2}{n_h n_{hi} n_{hij}}, \tag{22.1}$$

where

$$\begin{aligned} Z_h^2 &= S_h^2 - \frac{1}{N_h} \sum_i N_{hi} S_{hi}^2, \\ Z_{hi}^2 &= S_{hi}^2 - \frac{1}{N_{hi}} \sum_j N_{hij} S_{hij}^2. \end{aligned} \tag{22.2}$$

The Z_h^2 of (22.1) is not the same as that in (5.1) and this notation is maintained only for the sake of consistency. It will be assumed that Z_h^2 and Z_{hi}^2 are all positive. An expected cost function C_2 appropriate to (22) is

$$C_2 = C_o + \sum_{h=1}^{L} n_h c_h + \sum_{h=1}^{L} \frac{n_h}{N_h} \sum_{i=1}^{N_h} c_{hi} n_{hi} + \sum_{h=1}^{L} \frac{n_h}{N_h} \sum_{i=1}^{N_h} \frac{n_{hi}}{N_{hi}} \sum_{j=1}^{N_{hi}} n_{hij} c_{hij}, \tag{23}$$

where in stratum h, c_{hij} is the cost per third-stage unit in the j^{th} second-stage unit of the i^{th} first-stage unit.

The values of n_{hij}, n_{hi} and n_h which minimize cost or variance can be found from the following set of equations analogous to (2), viz.

$$\frac{N_h Z_h}{n_h \sqrt{c_h}} = \frac{N_h N_{hi} Z_{hi}}{n_h n_{hi} \sqrt{c_{hi}}} = \frac{N_h N_{hi} N_{hij} S_{hij}}{n_h n_{hi} n_{hij} \sqrt{c_{hij}}} = \lambda \text{ a constant.} \tag{24}$$

$$(h = 1,2,...L;\ i = 1,2,..N_h;\ j = 1,2,..N_{hi})$$

Thus we find

$$n_h = \frac{1}{\lambda} \cdot \frac{N_h Z_h}{\sqrt{c_h}} \tag{25}$$

$$n_{hi} = \sqrt{\frac{c_h}{c_{hi}}} \cdot \frac{N_{hi} Z_{hi}}{Z_h} \tag{26}$$

$$n_{hij} = \sqrt{\frac{c_{hi}}{c_{hij}}} \cdot \frac{N_{hij} S_{hij}}{Z_{hi}} \tag{27}$$

As demonstrated in the foregoing discussion, $\frac{1}{\lambda}$ can be determined either for the condition of fixed cost or fixed variance from (23) or (22.1), as the case may be, by substituting for n_h, n_{hi} and n_{hij} given at (25), (26) and (27) respectively.

In the multi-stage case when sampling is with equal probabilities and without replacement all the way, and with the same type of expected cost functions and linear estimator the equations for optima will be as in (24), only more extended, e.g.

$$\frac{N_h Z_h}{n_h\sqrt{c_h}} = \frac{N_h N_{hi} Z_{hi}}{n_h n_{hi}\sqrt{c_{hi}}} = \ldots = \frac{N_h N_{hi} N_{hij} N_{hijk} Z_{hijk}}{n_h n_{hi} n_{hij} n_{hijk}\sqrt{c_{hijk}}} = \ldots\ldots = \lambda \qquad (28)$$

The Z-functions can be identified from the variance function and the first two will be those given at (22.2). At the last stage of course, as can be seen from (22.1), only the variance internal to the unit will be involved so that the last expression in (28) will have a symbol of the type S_{hijklm} representing the square root of this internal variance for the unit in question.

Further for other sampling plans (e.g. sampling with unequal probabilities and with replacement at one or more intermediate stages) and with similar type of expected cost functions and linear estimators such as those considered above, formulas similar to (25), (26) and (27) will be obtained.

When non-response is taken into account, the optima for first-stage, second-stage sample sizes, etc. and the resampling rate or successive resampling rates for those units at the last stage (suffering from non-response) can be determined by exactly the same method. The structure of such sample designs may be described as multi-stage multi-phase. Unfortunately, the formulas are too long and unwieldly for presentation in an expository paper of this kind. These problems have also been considered recently by Mr. G. T. Foradori [4] in an unpublished work.

OPTIMUM ALLOCATION FOR SEVERAL ITEMS

This problem can be solved by the classical method of Lagrange but the equations obtained do not lead to algebraic solutions and can only be solved by iteration. Yates [5] has considered the problem for two variates. For more than two variates the computational work can become very tedious.

In one alternative approach, which is appealing because of its simplicity, the weighted sum of the relative variances of the variates of interest (corresponding to the items) is minimized for a fixed expected cost or the expected cost is minimized fixing the said weighted sum. The weights of course sum to one. The items are weighted according to their importance. The relative variance rather than the variance is considered because by so doing each component measure of precision is reduced to the same dimension and thus their linear compound is meaningful.

We shall investigate this problem in the light of the latter statement considering only stratified one-stage sampling with equal probabilities and without replacement. As will be seen subsequently, the problem for more than one stage of sampling, and with more involved probability systems, can be solved along the same lines.

Consider the estimation of the totals for r items (e.g. in a demographic survey, numbers employed, numbers unemployed, etc. in a region or country) using the best linear unbiased estimator for each.

The weighted sum of the relative variances will be

$$V = \sum_{t=1}^{r} w_t \frac{1}{X_t^2} \sum_{h=1}^{L} N_h^2 \frac{{}_tS_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right), \qquad (29)$$

where w_t is the weight assigned to the t^{th} item whose variance in the h^{th} stratum is ${}_tS_h^2$ and X_t is the total for the t^{th} item in the entire universe. The cost of the survey is given by

$$C = C_o + \sum_{h=1}^{L} c_h n_h \qquad (30)$$

The sumbols for which no explanation has been given here bear the same meaning as under Case (i), but restricted to one-stage sampling. (29) can be rewritten as

$$V' = V + \sum_{t=1}^{r} \frac{w_t}{X_t^2} \sum_{h=1}^{L} N_h\, {}_tS_h^2 = \sum_{h=1}^{L} \frac{N_h^2}{n_h} \sum_{t=1}^{r} w_t \frac{{}_tS_h^2}{X_t^2} \qquad (29.1)$$

We find

$$(C-C_o)V' \geq \left(\sum_{h=1}^{L} \frac{N_h}{\sqrt{n_h}}\sqrt{\sum_t w_t \frac{{}_tS_h^2}{X_t^2}} \cdot \sqrt{c_h n_h}\right)^2, \qquad (31)$$

and the minimum is attained when

$$\frac{\dfrac{N_h}{\sqrt{n_h}}\sqrt{\sum_t w_t \dfrac{{}_tS_h^2}{X_t^2}}}{\sqrt{c_h n_h}} = \lambda \text{ a constant.} \qquad (h = 1,2,\ldots L) \qquad (32)$$

From (32) we obtain

$$n_h = \frac{1}{\lambda} \cdot \frac{N_h}{\sqrt{c_h}}\sqrt{\sum_{t=1}^{r} w_t \frac{{}_tS_h^2}{X_t^2}}, (h=1,2,..L) \qquad (33)$$

Fixing the over-all relative variance in (29.1) at v_1 and eliminating λ between the resulting equation and (33) we obtain

$$n_h = \frac{N_h}{\sqrt{c_h}}\sqrt{\sum_t w_t \frac{{}_tS_h^2}{X_t^2}}\left(\sum_{h=1}^{L} c_h\; N_h\sqrt{\sum_t w_t \frac{{}_tS_h^2}{X_t^2}}\right) \Big/ \left(v_1 + \sum_t \frac{w_t}{X_t^2} \sum_h N_h\, {}_tS_h^2\right) \qquad (33.1)$$

Similarly when the total cost of the survey is fixed at c_1 we obtain

$$n_h = \frac{c_1 - C_o}{\sqrt{c_h}} \; N_h \sqrt{\sum_t w_t \frac{S^2_{th}}{X^2_t}} \Big/ \sum_h \sqrt{c_h} N_h \sqrt{\sum_t w_t \frac{S^2_{th}}{X^2_t}} . \tag{33.2}$$

In the light of the above discussion the interpretations to be made for the optima n_h given by (33.1) and (33.2) are readily apparent.

COMMENTS

Firstly, a question arises as to what should be done when one or more values such as the $Z^2_h = S^2_h - (\sum_i N_{hi} S^2_{hi}) / N_h$ are negative. There is no guarantee that they will always be positive since the underlying quadratic form in the variates concerned is not positive definite. Thus the square roots of the values concerned are no longer real but imaginary numbers. In any situation where the various parts of the variance function exhibit this refractory tendency the general method given above no longer holds, but, the basic procedure on which it rests, that is multiplying the cost and the variance function together, and determining the sample sizes which minimize the product, still holds. A discussion on this question is found in Sukhatme's book [2] on pp. 312-314, and will not be repeated here.

Secondly, the solutions for sample sizes presented here are formal. In practice integer values or such nearest values as are convenient to the survey organization will be used. Also they may usually be based on poor estimates of the unknown values involved. It can be demonstrated that "moderate" deviations from the optima do not increase the desired variance or alter the expected cost to an appreciable extent.

Finally, when sampling with unequal probabilities, as in Case (ii) above, the number of first-stage units may be fixed because variation in respect of the first stage is already controlled by the device. In such circumstances, the optima may also be determined as above for the remaining stage or stages. The only change is that those parts of the variance function and its corresponding cost function for example on the left hand sides of (16) and (17) respectively will be augmented by

$$-\sum_{h=1}^{L} \frac{U^2_h}{a} \quad \text{and} \quad -\sum_{h=1}^{L} c_h a$$

respectively, where a is the number of first-stage units per stratum selected on the basis of other considerations. The reasons behind such considerations may be of real importance in the conduct of sample surveys.

REFERENCES

[1] Neyman, J.. "Contribution to the Theory of Sampling Human Populations." J. Amer. Statist. Ass., 33 (1938) 101-116.

[2] Sukhatme, P. V. Sampling Theory of Surveys with Applications. New Delhi, India. The Indian Society of Agricultural Statistics. Ames, Iowa. The Iowa State College Press, 1953.

[3] Stuart, A. "A Simple Presentation of Optimum Sampling Results." J. Roy. Statist. Soc., B16 (1954), 239-241.

[4] Foradori, G. T. "Some Non-Response Sampling Theory for Two Stage Designs." Unpublished so far (1961).

[5] Yates, F. Sampling Methods for Censuses and Surveys. New York, Hafner Publishing Company, (1953, 2nd edition), 338-342.