J. R. Statist. Soc. A,
(1976), 139, *Part* 1, p. 80

Optima and Proxima in Linear Sample Designs

By Leslie Kish†

University of Michigan

Summary

The distinct problems of allocating $\sum m_i$ sampling units are stated jointly as minimizing $(\sum V_i^2/m_i)(\sum m_i)$, when either term is fixed, where $\text{var}(\bar{y}) = \sum V_i^2/m_i + V_0$ and $\text{cost}(\bar{y}) = \sum c_i m_i + C_0$. The V_i^2/m_i and the $c_i m_i$ are variance and cost components (strata, stages, phases, etc.); these are affected by the allocation of the m_i, but the V_0 and C_0 are not. The Lagrange identity yields the general relative loss function: $1+L = (\sum U_i k_i)(\sum U_i/k_i)$, where L is the relative loss, due to relative departures $k_i \propto m_i^*/m_i$ from the optimal $L = 0$, where the $m_i^* \propto V_i/\sqrt{c_i}$, and the $U_i = V_i\sqrt{c_i}/\sum V_i\sqrt{c_i}$ are relative measures of the components. Solutions are applied to the distinct problems of sample designs. Tables are given of the loss L for useful models of U_i and k_i. The method also leads to useful compromises among the conflicting aims of multipurpose samples, and to measures of relative losses for those aims, e.g. in the conflict between domains and overall means.

Keywords: OPTIMAL ALLOCATION; MULTIPURPOSE ALLOCATION; SAMPLE DESIGNS; OPTIMAL SAMPLING; EFFICIENT DESIGNS; LINEAR DESIGNS; STRATIFIED ALLOCATION; RELATIVE PRECISION; LAGRANGE IDENTITY; DOMAIN ALLOCATION

1. Basic Definitions and Aims

In the literature of survey sampling diverse problems of optimal allocation are treated separately. Yet they can usefully be viewed as distinct examples of the same simple expressions for the total variance and cost of the sample statistic $\bar{y}$:

$$\text{var}(\bar{y}) = V + V_0 = \sum V_i^2/m_i + V_0 \tag{1.1}$$

and

$$\text{cost}(\bar{y}) = C + C_0 = \sum c_i m_i + C_0. \tag{1.2}$$

These linear forms occur in stratified, multistage and multiphase sampling, and other related techniques. Of several applications in Section 7, consider two specific examples. (a) For a stratified sample of elements the variance of the mean $\bar{y} = \sum W_i \bar{y}_i$ is:

$$\text{var}(\bar{y}) = \sum (W_i S_i)^2/m_i - \sum (W_i S_i)^2/M_i,$$

where m_i, M_i, W_i and S_i^2 are respectively the sample and population sizes, the weights and element variances in the ith stratum. The first term, V, depends on the allocation of the m_i; the second, V_0 does not. (b) For two-stage random subselection of b from B elements within each of a random selections from A clusters, the variance of the mean is:

$$\text{var}(\bar{y}) = (1-a/A)\,S_a^2/a + (1-b/B)\,S_b^2/ab = S_u^2/a + S_b^2/ab - S_a^2/A.$$

Here $S_u^2 = S_a^2 - S_b^2/B$; V comprises the first two terms, with $m_1 = a$ and $m_2 = ab$; the last term $-S_a^2/A = V_0$ does not depend on the m_i. The cost is $c_a a + c_b ab + C_0$.

Definitions and restrictions seem desirable here.

(1) The *statistic* $\bar{y}$ denotes an estimate of a mean or of an aggregate. Possible extensions to other estimates are not attempted here.

(2) The ith component of the variance, V_i^2/m_i, denotes a constant V_i^2 in the design, a unit variance, divided by the number m_i of sampling units for that component. We prefer V_i^2 to V_i to denote unit variances that are commonly defined with squared values.

† Research Scientist, Institute for Social Research and Professor at the University of Michigan.

(3) The ith component of cost, $c_i m_i$, denotes the unit cost c_i multiplied by the same number m_i of units as in (2).

(4) *Components* may refer to strata or stages or phases of sampling; generality is the essence of our approach. Components here represent additive sources of variation and cost.

(5) The constants V_i^2 and c_i are parameters for which values are assumed or guessed for numerical solutions of allocation problems. We take $V_i \geqslant 0$ and $\sqrt{c_i} \geqslant 0$ (hence V_i^2 and c_i) for allocating the m_i. For non-triviality two pairs at least of the V_i and c_i should be positive. Negative values of V_i^2 may be encountered, as with S_u^2 above; we then redefine the problem to facilitate a practical solution; for an example see Section 7.3.

(6) The constants V_0 and C_0 do not affect optimal allocations of the m_i; their effects on losses in proximal allocation are shown in Section 3. C_0 is non-negative in practice, but V_0 is often negative, as above.

(7) For practical values of m_i we want positive integers. Also $0 < m_i \leqslant M_i$, where M_i denotes the number of units in the population for the component; and $m_i \geqslant 2$ for computing variance components. Frequently, allocation formulae yield some optimal values of $m_i^* > M_i$; when these are reset to $m_i = M_i$ the other optimal values of $m_i^* < M_i$ can be recomputed with (5.4) and (5.5).

(8) It would seem more realistic to guess distributions for V_i^2 and c_i, rather than single values, and a Bayesian treatment of design will probably be worth while. But that is beyond our scope here, and I dread a complex procedure out of the reach of survey practitioners. Furthermore, its relative losses would probably not differ much from ours, because losses are insensitive to moderate departures from the guesses.

(9) In some applications, especially for some stratified samples, differences between the c_i are disregarded. Hence, the cost constraint becomes $C/c = m = \sum m_i$. Then the $\sqrt{c_i}$ should be omitted from the allocation formulae. Instead of C_0 use C_0/c, where c is a common (average) unit cost.

(10) This last point calls attention to the dimensional (unit) homogeneity of all the formulae.

To find optimal values $m_i^* \propto V_i/\sqrt{c_i}$ for the m_i we minimize the product

$$VC = (\sum V_i^2/m_i)(\sum c_i m_i), \tag{1.3}$$

when either V or C is fixed at V_f or C_f. This results in the same optimal values as

$$\text{var}(\bar{y}) \times \text{cost}(\bar{y}) = (V+V_0)(C+C_0),$$

because in $(V+V_0)C_f$ or in $V_f(C+C_0)$ the second terms are unaffected by optimal allocation; their effects on proxima are more easily treated separately (Section 3). To use the product VC rather than some other function seems reasonable: an increase (or decrease) in cost by some factor should be equivalent to a decrease (or increase) in variance by the same factor. The product form leads directly to expressions for loss functions $(1+L)$ and relative losses (L) that are our goals here. For brevity, I use "loss" for L that represents relative increase of variance or cost, without limits.

Our principal aim goes beyond optimization of linear forms, to a simple and coherent treatment of their *proximization*. We provide convenient forms, in terms of useful parameters, for *relative losses* incurred by *proxima* achieved with *proximal* allocations.

To *proximize* has the same flavour as Herbert Simon's "to satisfice," but I prefer here the former's neutrality, flexibility and its several forms resembling those of optimize. With these neologisms I want to emphasize a statistical approach, as complementary to the mathematical concept of optimization. Since we cannot attain full optima with designs based on guessed parameters, it is equally important to guess relative losses due to expected sets of departures from optimal allocations. Here we view the optimum as the limit 1 for the loss function $(1+L)$ as L approaches 0, and how that approach is made. Two designs, one that loses 2 per cent and the other 50, are both non-optimal by strict "optimist" standards; but a

"proximist" would usually class a 2 per cent loss with the optimal, to distinguish both from larger losses like 50 per cent.

This illustrates that for statisticians "The perfect is the enemy of the good" (proximized from Voltaire). Conflict appears frequently; optimization for one convenient variable often usurps the place of proximization for multipurpose allocation. Proximal methods are seen to be particularly adaptable to multipurpose allocation in Sections 6 and 7.6, and fulfil our second aim.

Further, we also present Section 5, a compact, simple and general formulation of optimal allocations for diverse sampling methods. Instead of solving each separately, we merely substitute appropriate symbols for the optimal values $m_i^* \propto V_i/\sqrt{c_i}$. This is obtained with the simple Cauchy inequality. This unified and simple treatment has heuristic and pedagogic merit. Applications in Section 7 cover the diversity of sampling methods. Sections 2 and 3 develop methods of proximization, and Section 4 contains convenient tables for relative losses L.

2. General Formulation

Our principal result (2.3) expresses the relative loss (L) in two parameters: U_i, the relative "sizes" of the components; and $k_i \propto m_i^*/m_i$, the relative departures of the sample sizes m_i from optimal allocations m_i^*. First (2.1), the product VC to be minimized is divided by $(\sum V_i\sqrt{c_i})^2$; this ratio will be shown to have minimal (optimal) value of 1. It expresses the *relative loss* L for any allocation of the m_i (>0), by compensating for the units of measurement of the V_i^2 and c_i. Next (2.2), the m_i are stated in terms of relative departures $k_i \propto m_i^*/m_i$ from their optimal values m_i^*; these will be shown to be $m_i^* \propto V_i/\sqrt{c_i}$. Hence we substitute $m_i \propto V_i/\sqrt{(c_i)}\,k_i$ to obtain (2.2); the factors of proportionality cancel. Finally, (2.3) for generality and brevity we substitute the relative "sizes" $U_i = V_i\sqrt{c_i}/\sum V_i\sqrt{c_i}$.

$$1+L = VC/(\sum V_i\sqrt{c_i})^2 = (\sum V_i^2/m_i)(\sum c_i m_i)/(\sum V_i\sqrt{c_i})^2, \tag{2.1}$$

$$= \{\sum V_i\sqrt{(c_i)}\,k_i\}(\sum V_i\sqrt{c_i}/k_i)/(\sum V_i\sqrt{c_i})^2, \tag{2.2}$$

$$= (\sum U_i k_i)(\sum U_i/k_i), \tag{2.3}$$

where $U_i = V_i\sqrt{c_i}/\sum V_i\sqrt{c_i}$ and $k_i \propto V_i/\sqrt{(c_i)}\,m_i$. We take the k_i and U_i to be positive and finite. We have $\sum U_i = 1$; and we may also use any convenient $U_i' \propto U_i$, if we divide by $\sum U_i'$. Note that we need only the relative values of k_i, and we can use Ak_i, with A any positive and finite constant. Furthermore, the form of (2.3) shows that the k_i may be replaced by their reciprocals; they may refer to ratios of oversampling, as well as undersampling. With this flexibility we can use $\min(k_i) = 1$, as we do in Table 1 for convenience.

The minimal value of 1 for (2.3) is obtained with all $k_i = k^*$ equal. This may appear obvious, or seen with the Lagrange identity in (5.1).

Examples may be useful here.

(a) Consider the variance of the mean $\sum W_i\bar{y}_i$ for two strata where $W_1 = 0{\cdot}2$, $W_2 = 0{\cdot}8$, $S_1^2 = S_2^2 = S^2$ and $c_1 = c_2 = c$. Then $U_i = W_i \propto V_i$, and $U_1 : U_2 = 1 : 4$. This implies (5.3) that optimal allocation of sample sizes should be in the ratio of stratum sizes W_i, hence $m_2 = 4m_1$. If samples of equal sizes, $m_1 = m_2$, are taken, this implies a departure factor of 4; we can use simply $k_1 = 1$ and $k_2 = 4$. The consequent relative loss L would be given by (2.3) as $1+L = (0{\cdot}2\times 1+0{\cdot}8\times\frac{1}{4})\,(0{\cdot}2\times 1+0{\cdot}8\times 4) = 1{\cdot}360$.

(b) To illustrate the effect of the U_i on the loss L: suppose now $S_1^2 = 4S_2^2$, and $c_1 = 4c_2$. Since $S_1^2/c_1 = S_2^2/c_2$, optimal allocation is still $1 : 4$. But now $U_i = W_i S_i\sqrt{c_i}/\sum W_i S_i\sqrt{C_i}$, hence $U_1 = U_2 = 0{\cdot}5$. Therefore the relative loss L from equal sample sizes now would be given by $1+L = (0{\cdot}5\times 1+0{\cdot}5\times\frac{1}{4})\,(0{\cdot}5\times 1+0{\cdot}5\times 4) = 1{\cdot}5625$.

(c) To illustrate a conflict in allocation: suppose that, as in (a), $S_1^2 = S_2^2$ and $c_1 = c_2$, but that now we want to minimize the variance of the difference of means $(\bar{y}_1-\bar{y}_2)$. Now $U_1 = U_2 = 0{\cdot}5$. Optimal allocation is at $m_1 = m_2$. Departure from this in the ratio $1 : 4$ to satisfy (a) would result in $1+L = (0{\cdot}5\times 1+0{\cdot}5\times 4)\,(0{\cdot}5\times 1+0{\cdot}5\times\frac{1}{4}) = 1{\cdot}5625$.

Note that these answers can also be found in Table 1 in column $K = 4$ for relative departures. The size of U of one component is 0·2 for (a), and 0·5 for both (b) and (c), in the top two rows. Results for (a) and (c) illustrate common conflicts between totals and domains, treated in Section 7.6, and in Table 5(A).

The weights U_i are convenient for design; based on population parameters, we may call them *population weights*. However, when dealing with sample results it may be more convenient to use *sample weights*, based on sample sizes: $u_i = U_i/k_i$. Then (2.3) may be written as

$$1+L = (\sum u_i k_1^2)(\sum u_i), \tag{2.4}$$

$$= 1+(\sum u_i k_1^2/\sum u_i - \bar{k}^2)/\bar{k}^2 = 1+C_k^2, \tag{2.5}$$

$$= 1+\sum u_i (k_i/\bar{k}-1)^2/\sum u_i, \tag{2.6}$$

$$= 1+\sum (k_i/\bar{k}-1)^2 m_i c_i/\sum m_i c_i. \tag{2.7}$$

C_k^2 is the relative variance (relvariance) of the k_i with sample weights u_i around their mean $\bar{k} = \sum u_i k_i/\sum u_i = 1/\sum u_i$. Here larger $k_i > 1$ represent larger weights to compensate for undersampling proportionately to their reciprocals. The $u_i = U_i/k_i$ are proportional to $c_i m_i$ because the $U_i = V_i\sqrt{c_i}/\sum V_i\sqrt{c_i} \propto c_i m_i^* \propto c_i m_i k_i$.

3. On Proximal Allocation

Extreme departures from optimal values of m_i^* can result in large relative losses measured in either cost or variance. However, small or even moderate departures from the optimal m_i^* lead only to negligible or small relative losses. These vague precepts of practising statisticians are given formal and practical expressions (2.4)–(2.7) in terms of the relative loss (L) compared to optimal allocation.

Departures from optimal allocation have several causes. In actual surveys some departures are unavoidable because the true and exact values of V_i^2 and c_i are not available. Second, we may further depart from indicated optimal values m_i^* to convenient proximal integers, or to convenient sampling fractions. Third, surveys generally have several or many purposes, for which the optimal allocations are different. Fourth, the nature of the sampling frame and of the data collection may force departures from optimal selection probabilities; sometimes we are forced to accept unequal selection probabilities when equal probabilities would be close to optimal. Fifth, mistakes in design may be added to the list of good reasons. And sixth, departures from computed values m_i^* may be forced by the constraints $m_i \leqslant M_i$ and $m_i \geqslant 2$, as noted above.

Thus on many occasions we find it useful to have simple formulations for the relative losses brought about by a set of departures from the optima. We can give useful approximations in terms of the factors k_i of relative departures, and of the weights, for different components i of the sample.

When the frequencies for the k_i are given or estimated in *sample proportions* u_i, then (2.5)–(2.7) yield readily the loss L in terms of the *relvariance* C_k^2 of the k_i values. One of these formulae may be most convenient for judging the losses from actual sample results.

However, for comparing designs of planned samples the frequencies may be more conveniently stated in terms of the *population weights* U_i. Formula (2.3) can be readily computed for moderate numbers of components. Furthermore, the simple models of Table 1 can often give instant answers for approximate distributions. I have often found these answers close and adequate for planning designs.

Computations of the relative loss L in our formulae and tables take account of the factors V and C in the minimized function VC, but they neglect the constants V_0 and C_0 in the total variance and cost (1.1 and 1.2) However, this neglect may be corrected with translations of L into L' that does take into account the constant factors V_0 and C_0. If $V_{\min}$ is the optimal V

for fixed C_f then the ratio of the attained proximal variance to the optimal variance is

$$\frac{(l+L)V_{\min}+V_0}{V_{\min}+V_0} = 1+L/(1+V_0/V_{\min}) = 1+L'. \tag{3.1}$$

Thus the adjusted *actual relative loss* L' differs from that indicated by L; since V_0 is often negative, L' can be somewhat greater than L. For a $C_{\min}$ found for a fixed V_f, the adjusted relative loss L' may be somewhat less than L due to a positive C_0 in

$$L' = L/(1+C_0/C_{\min}). \tag{3.2}$$

4. Tables of Losses for Model Distributions

For a variety of simple models we can give instant answers about expected losses. Actual population distributions can usually be matched against one of these models so as to provide useful approximations of the expected losses.

The losses are given in terms of departures k_i from optimal allocations for the relative weights U_i in the models, and the k_i range from $\min(k_i) = 1$ to $\max(k_i) = K$. The simplest model consists of two components U and $(1-U)$, where the relative departures from optimal sample sizes are in the ratio $k_1 : k_2 = 1 : K$. The loss for two components may be expressed (7.4) as

$$L = U(1-U)(K-1)^2/K. \tag{4.1}$$

The dichotomous models represent maximal losses for ranges of departures fixed at 1 to K. Thus losses for large values of K are much greater in the top three rows of Table 1 than further down where five other models are shown.

The five models represent diverse frequency distributions for the population weights U_i; and for each model both discrete and continuous versions are shown. In the discrete versions the relative departures k_i take K integral values from 1 to K, and the relative weights U_i are concentrated at those values. In continuous versions the departures k_i and relative weights U_i vary continuously from 1 to K. Frequencies are divided by their sums to produce relative frequencies U_i.

Note that the loss L is both very small and uniform for all models for small K; for $K = 1{\cdot}3$, $L_d = 0{\cdot}017$ and $L_c = 0{\cdot}006$; for $K = 1{\cdot}5$, $L_d = 0{\cdot}04$ and $L_c = 0{\cdot}014$. (Note that for L_d the k_i take only two values: 1 and $K = 1{\cdot}3$ or $1{\cdot}5$). From $K = 2$ to about $K = 5$ the losses are moderate and fairly similar for the five models. The L_c are lower than the L_d, though in an irregular ratio. Below $K = 10$ we can make fairly good guesses about L just from the range 1 to K, without knowing much about the U_i—if this is not dichotomous or U-shaped.

However, beyond $K = 10$ the losses L increase and diverge. Three of the models show rather similar losses, but for the model $U_i \propto 1/k_i$ the losses are much larger. And this model, may often resemble actual frequencies. The fifth has much lower losses, but it is not realistic, I think.

From the models one can also make conjectures about actual distributions that differ somewhat from them. For example, a rectangular distribution for integral values of k_i from 1 to 5 has $L_d = 0{\cdot}370$; more than five values evenly spaced within the same range to 1 to 5 would have a loss between that value and the continuous loss $L_c = 0{\cdot}207$. On the other hand, for only three values of k_i from 1 to 5 and $U_i = \frac{1}{3}$, the loss (actually 0·533) is above 0·370, but below the dichotomous value of 0·800 in Table 1.

When *sample weights* $u_i = U_i/k_i$ seem more convenient, the relative loss L may be estimated by the relvariance C_k^2 of the k_i, with weights u_i (2.7). Formulae and tables can be constructed for such relvariances, if we begin with the means M and variances σ^2 of convenient distributions from 0 to 1 (Kish, 1965, p. 262). To obtain the relvariances C_k^2, those variances are multiplied by the new $(\text{range})^2 = (K-1)^2$ and divided by the new $(\text{mean})^2 = \{M(K-1)+1\}^2$; thus $C_k^2 = \sigma^2(K-1)^2/\{M(K-1)+1\}^2$. Table 3 notes six useful examples.

TABLE 1

Relative losses (L) for six models of population weights (U_i); for discrete (L_d) and continuous (L_c) weights: for relative departures (k_i) in the range from 1 to K

Models	*K*	1·3	1·5	2	3	4	5	10	20	50	100	500	1,000
Dichotomous $U(1-U)$													
(0·5) (0·5)		0·017	0·042	0·125	0·333	0·562	0·800	2·025	4·512	12·005	24·50	124·5	249·5
(0·2) (0·8)		0·011	0·027	0·080	0·213	0·360	0·512	1·296	2·888	7·683	15·68	79·7	159·7
(0·1) (0·9)		0·006	0·015	0·045	0·120	0·202	0·288	0·729	1·624	4·322	8·82	44·8	89·8
Rectangular	L_d	0·017*	0·042*	0·125*	0·222	0·302	0·370	0·611	0·889	1·295	1·620	2·403	2·746
$U_i \propto 1/K$	L_c	0·006	0·014	0·040	0·099	0·155	0·207	0·407	0·656	1·036	1·349	2·120	2·461
Linear decrease	L_d	0·017*	0·040*	0·111*	0·203	0·283	0·353	0·616	0·940	1·437	1·917	2·879	3·333
$U_i \propto K+1-k_i$	L_c	0·006	0·014	0·040	0·097	0·153	0·205	0·409	0·680	1·127	1·514	2·507	2·956
Hyperbolic decrease	L_d	0·017*	0·040*	0·111*	0·215	0·312	0·404	0·807	1·466	3·014	5·076	16·802	28·342
$U_i \propto 1/k_i$	L_c	0·006	0·014	0·041	0·103	0·171	0·235	0·528	1·011	2·138	3·621	11·998	19·915
Quadratic decrease	L_d	0·016*	0·036*	0·080*	0·150	0·211	0·264	0·460	0·696	1·048	1·333	2·026	2·331
$U_i \propto 1/k_i^2$	L_c	0·006	0·014	0·040	0·099	0·155	0·207	0·407	0·656	1·036	1·349	2·120	2·461
Linear increase	L_d	0·017*	0·040*	0·111*	0·167	0·200	0·222	0·273	0·302	0·320	0·327	0·330	0·333
$U_i \propto k_i$	L_c	0·006	0·013	0·037	0·088	0·120	0·148	0·223	0·273	0·308	0·320	0·331	0·332

Dichotomous $1+L = 1+U(1-U)(K-1)^2/K$
Discrete $1+L_d = (\sum U_i k_i)(\sum U_i/k_i)$, with $k_i = i = 1, 2, 3, \ldots, K$
Continuous $1+L_c = \int Uk.dk. \int (U/k)\,dk$, with $1 \leqslant k \leqslant K$.
Only two values, 1 and K, were used for L_d for $K = 1{\cdot}3$, 1·5 and 2

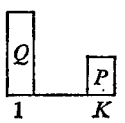
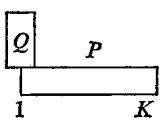
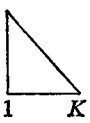
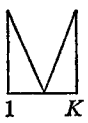
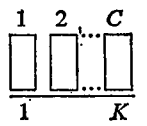
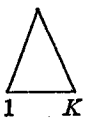
TABLE 2

Formulae for loss functions $(1+L)$ *due to relative departures* k_i *for five models, for both discrete and continuous*

Models	$U_i \propto 1/K$	$U_i \propto K+1-k_i$	$U_i \propto 1/k_i$	$U_i \propto 1/k_i^2$	$U_i \propto k_i$
Discrete $1+L_d$	$\frac{K+1}{2K}\sum\frac{1}{i}$	$\frac{2(K+2)}{3}\left(\frac{1}{K}\sum\frac{1}{i}-\frac{1}{K+1}\right)$	$\frac{K\sum 1/i^2}{(\sum 1/i)^2}$	$\frac{\sum 1/i.\sum 1/i^3}{(\sum 1/i^2)^2}$	$1+\frac{K-1}{3(K+1)}$
Continuous $1+L_c$	$\frac{K+1}{2(K-1)}\ln K$	$\frac{2(K+2)^2-3}{3(K+1)(K-1)}\left(\ln K-\frac{K-1}{K+1}\right)$	$\frac{(K-1)^2}{K(\ln K)^2}$	$\frac{K+1}{2(K-1)}\ln K$	$1+\frac{(K-1)^2}{3(K+1)^2}$

TABLE 3

Losses L *for six models of sample weights* $u_i = U_i/k_i$*; the departures* $k_i \geqslant 1$ *from* 1 *to* K *represent compensations for undersampling*

1	2	3	4	5	6
					
$\frac{Q}{P}\left(\frac{K-1}{K+Q/P}\right)^2$	$\frac{1+3Q}{3P}\left(\frac{K-1}{K+1+2Q/P}\right)^2$	$\frac{1}{2}\left(\frac{K-1}{K+2}\right)^2$	$\frac{1}{2}\left(\frac{K-1}{K+1}\right)^2$	$\frac{1+2/C}{3}\left(\frac{K-1}{K+1}\right)^2$	$\frac{1}{6}\left(\frac{K-1}{K+1}\right)^2$

Losses (L) can get large for models 1 and 2 when both K and Q are great: for $K = 20$ and $Q = 10P$ the L is 4·011 for 1 and 2·219 for 2. For the other models losses remain moderate.

5. ON OPTIMAL ALLOCATION

The Lagrange identity is a basic tool of great utility, and it may be stated here simply. Assume x_i and y_i $(i = 1, 2, \ldots, n)$ finite and real; but here we need only non-negative values. Then

$$\begin{aligned}(\sum x_i^2)(\sum y_i^2) &= \sum x_i^2 y_i^2 + \sum_{i\neq j} x_i^2 y_j^2 \\ &= \sum x_i^2 y_i^2 + \sum_{i\neq j} x_i y_i x_j y_j + \sum_{i\neq j} x_i^2 y_j^2 - \sum_{i\neq j} x_i y_i x_j y_j \\ &= (\sum x_i y_i)^2 + \sum_{i<j}(x_i y_j - x_j y_i)^2. \end{aligned} \tag{5.1}$$

The second term has a minimum of 0, when $y_i = Fx_i$, F constant. The first term alone is the lower bound of the Cauchy–Schwartz inequality. If we take in (5.1) $x_i = \sqrt{(U_i/k_i)}$ and $y_i = \sqrt{(U_i k_i)} = k_i x_i$ we now rewrite (2.3) as

$$\begin{aligned}1+L &= (\sum U_i k_i)(\sum U_i/k_i) = \sum y_i^2 \sum x_i^2 \\ &= 1 + \sum_{i<j} \frac{U_i U_j}{k_i k_j}(k_i - k_j)^2, \end{aligned} \tag{5.1'}$$

with $(\sum x_i y_i)^2 = (\sum U_i)^2 = 1$. The minimal value is 1, when the second term is 0, because all k_i are equal.

Now let $x_i = \sqrt{(V_i^2/m_i)}$ and $y_i = \sqrt{(c_i m_i)}$, with V_i^2 and c_i as assumed parameters and m_i as variables (all $\geqslant 0$). The minimal value of

$$VC = (\sum V_i^2/m_i)(\sum c_i m_i) \geqslant (\sum V_i \sqrt{c_i})^2, \tag{5.2}$$

a Cauchy–Schwartz inequality, is obtained when $\sqrt{(c_i m_i^*)} = F\sqrt{(V_i^2/m_i^*)}$. Then

$$m_i^* = FV_i/\sqrt{c_i} \tag{5.3}$$

are the optimal values of the m_i^* that obtain the

$$\text{minimal } VC = (\sum V_i \sqrt{c_i})^2. \tag{5.3'}$$

The constant F can be determined from either C_f or V_f fixed. With $C_f = \sum c_i m_i^* = F \sum V_i \sqrt{c_i}$ one uses $F = C_f/\sum V_i \sqrt{c_i}$. For $V_f = \sum V_i^2/m_i^*$ note that $V_i \sqrt{c_i} = FV_i^2/m_i^*$ and $\sum V_i \sqrt{c_i} = FV_f$; hence $F = (\sum V_i \sqrt{c_i})/V_f$.

Further $VC/(\sum V_i \sqrt{c_i})^2 = 1$ yields either $V_{\min}$ for C_f or $C_{\min}$ for V_f. The minimal value of 1 is viewed as the limiting value of $(1+L)$, as the second term approaches zero and as the m_i approach the optimal values m_i^*. The relative size of the second term is seen, when divided by the minimal first term $(\sum V_i \sqrt{c_i})^2$, as the relative loss L, due to proximization.

This formulation of the relative loss is the principal advantage of the Lagrange identity over the Cauchy–Schwartz inequality, which has been used for optimizing stratified samples (Stuart, 1954; Sukhatme, 1970); also over the method of Lagrange multipliers, generally used in sampling literature as $F(m_i) = \sum V_i^2/m_i + \lambda(\sum c_i m_i - C)$, to obtain the same optima. This method, however, can deal with more complex functions, such as $C = c_0 \sqrt{m_1} + c_1 m_1 + c_2 m_2$.

The optimal formulae may give unacceptable answers when high values of $V_i/\sqrt{c_i}$ result in $m_i^* > M_i$, the population of units of the ith component. For example, this occurs often in strata with high values of S_i; also when subsamples from clusters must be confined to the cluster size (Section 7.3). For such components set $m_i^* = M_i$; we may also do this arbitrarily for components when $m_i^* > M_i/2$, for example. However, changing only those m_i^* to $m_i' = M_i$ would cause the constraint V_f or C_f to be missed. To re-establish the constraint we must recompute for the other values of m_i^* new, higher values m_i' proportional to them.

Consider the components V_i^2/m_i and $c_i m_i$ divided into two sets: the complete (q), where $m_i' = M_i$; and the partial (p), where $m_i' < M_i$; denote the sets as $\sum_i = \sum_q + \sum_p$. The complete components $\sum_q$ are transferred to the constant factors V_0 and C_0: where $m_i' = M_i$ the m_i' cease to affect the allocation, because components of the variance with the form $V_i^2(1/m_i' - 1/M_i)$ vanish for $m_i' = M_i$. Thus

$$\text{var}(\bar{y}) = \sum_i \frac{V_i^2}{m_i'} + V_0 = \sum_p \frac{V_i^2}{m_i'} + V_0 + \sum_q V_i^2/M_i \tag{5.4}$$

and

$$\text{cost}(\bar{y}) = \sum_i c_i m_i' + C_0 = \sum_p c_i m_i' + C_0 + \sum_q c_i M_i. \tag{5.5}$$

Then we may solve (5.2) and (5.3) with only p components in one of the first terms above and in $\sum V_i \sqrt{c_i}$. We compute the residual constant either as $V_p = \text{var}(\bar{y}) - V_0 - \sum_q V_i^2/M_i$, or as $C_p = \text{cost}(\bar{y}) - C_0 - \sum_q c_i M_i$.

For convenience, since the m_i^* are already in the right proportions, instead of recomputing (5.2) and (5.3) we may merely increase the m_i^* proportionately to the new values m_i' in the subset $\sum_p$. Thus: for fixed partial cost C_p, increase the m_i^* to $m_i' = m_i^*(C_p/\sum_p c_i m_i^*)$, or for fixed partial variance V_p, increase the m_i^* to $m_i' = m_i^*(\sum_p V_i^2/m_i^*)/V_p$.

Note also that optimal values $V_{\min}$ and $C_{\min}$ obtained with (5.2) without these adjustments should be recomputed to obey the constraints $m_i \leqslant M_i$. These enforced departures from the optimal m_i^* will tend to increase the V or C that can be considered as attainable under the restraints. This will affect the relative loss, as V_0 and C_0 are treated at the end of Section 3.

6. MULTIPURPOSE ALLOCATION

Sample surveys are typically multipurpose in nature, and it seems imperative to extend the methods of allocation to multipurpose designs. For lack of these methods univariate allocation dominates our literature and theory of sampling; practical work is also affected, but less often. The methods for optimization and proximization developed here seem particularly adaptable to multipurpose design. The general form $\sum V_i^2/m_i$ for variances can serve well the many purposes of a sample survey; for the gth purpose the variance will be denoted by $\sum_i V_{gi}^2/m_i$.

The many purposes of a single survey may have several sources: (1) A single variable may result in several statistics; e.g. the mean and median of incomes can benefit from different allocations (Kish, 1961). (2) Most surveys obtain results for several variables on a single subject. (3) Furthermore, some surveys are *multisubject* in character; e.g. with economic, demographic, social variables. (4) Results for subclasses and for their comparisons may be as important as results based on the entire sample. Designs for subclasses often point to different designs and allocations than those for the entire sample. (5) The common but neglected conflict between designs for comparisons between domain means and for the combined mean for the entire sample is developed in Section 7.6.

Suppose a sample is allocated optimally for variate $\bar{y}'$ with m_i' proportional to $V_i'/\sqrt{c_i'}$, but optimal allocation for another variate $\bar{y}$ would be $m_i \propto V_i/\sqrt{c_i}$. The loss incurred for $\bar{y}$ can be measured with the departures $k_i = m_i/m_i' = (V_i/V_i')(\sqrt{c_i'}/\sqrt{c_i})$ and with weights $U_i = V_i\sqrt{c_i}/\sum V_i\sqrt{c_i}$ in formula (2.3). We are mostly concerned with allocation of the m_i within one survey sample, so that $\sqrt{c_i'}/\sqrt{c_i} = 1$. Then the loss function for $\bar{y}$ due to optimization for $\bar{y}'$ may be represented by

$$1+L(m_i') = (\sum V_i^2\sqrt{c_i}/V_i')(\sum V_i'\sqrt{c_i})/(\sum V_i\sqrt{c_i})^2$$

$$= \sum\left(\frac{V_i\sqrt{c_i}}{\sum V_i\sqrt{c_i}}\right)^2 \Big/ \left(\frac{V_i'\sqrt{c_i}}{\sum V_i'\sqrt{c_i}}\right). \tag{6.1}$$

This may be regarded as the relvariance of $k_i = V_i/V_i'$ with weights $u_i = V_i'\sqrt{c_i}$ (2.7). Often the cost factors are constant or disregarded, and (6.1) has a particularly simple form

$$1+L(m_i') = \sum (V_i/\sum V_i)^2/(V_i'/\sum V_i'). \tag{6.1'}$$

If the m_i allocated for one survey with c_i are used for another with $c_i \neq c_i'$, then we rewrite (6.1), with $V_i'\sqrt{c_i}/\sqrt{c_i'}$ in place of V_i', as

$$1+L(m_i') = (\sum V_i^2\sqrt{c_i'}/V_i')(\sum V_i'\sqrt{c_i}/\sqrt{c_i'})/(\sum V_i\sqrt{c_i})^2. \tag{6.2}$$

Now consider a loss function for several variates indexed with $(g = 1,2,3,\ldots)$. The loss function, for a fixed cost $C_f = \sum c_i m_i$, may be expressed for each as $1+L_g = (\sum V_{gi}^2/m_i)/V_{g\min}$, where the denominator denotes the minimal variance attainable and computed for the gth variate. Assign the weights I_g $(\sum I_g = 1)$ to denote *the relative importance* of the lost precision of the gth variate. Then consider the total expected loss as a linear function of the quadratic loss functions (for a fixed set of m_i) of the variances

$$1+L(m_i) = \sum_g I_g(1+L_g) = 1+\sum_g I_g L_g(m_i) = \sum_g I_g\left\{\left(\sum_i V_{gi}^2/m_i\right)\Big/ V_{g\min}\right\}$$

$$= \sum_i \frac{1}{m_i}\sum_g \frac{I_g V_{gi}^2}{V_{g\min}} = \sum_i \frac{Z_i^2}{m_i}, \tag{6.3}$$

where $Z_i^2 = \sum_g I_g V_{gi}^2/V_{g\min}$. Changing the order of summation permits defining this ith component that can be computed. For the multipurpose joint allocation we may compute

(5.3) the

$$\text{optimal } m_i^{**} = \frac{Z_i}{\sqrt{c_i}} \frac{C_f}{\sum Z_i \sqrt{c_i}} \tag{6.4}$$

and

$$1+L(m_i^{**}) = V_{\min} = (\sum Z_i \sqrt{c_i})^2 / C_f. \tag{6.5}$$

From the multipurpose optimal allocations m_i^{**} we may compute the loss function $1+L_g(m_i^{**})$ for the gth variate considered separately. For each of these we can use (6.1) with $V_i = V_{gi}$, $V_i' = Z_i$, $k_{gi} = V_{gi}/Z_i$ and $U_{gi} = V_{gi}\sqrt{c_i}/\sum V_{gi}\sqrt{c_i}$. These may be averaged with the weights I_g to obtain the joint loss function (6.3) of $1+L(m_i^{**})$ with the multipurpose optimal allocations m_i^{**}.

This however may be obtained more directly from (6.4) or (6.5). Thus

$$\begin{aligned} 1+L(m_i^{**}) &= \sum_i \frac{Z_i^2}{m_i^{**}} = \left(\sum_i Z_i \sqrt{c_i}\right)^2 \Big/ C_f \\ &= \frac{1}{C_f}\left\{\sum_i \sqrt{\left(\sum_g \frac{I_g V_{gi}^2 c_i}{V_{g\min}}\right)}\right\}^2. \end{aligned} \tag{6.6}$$

When we accept (from (5.2) $V_{g\min} = (\sum V_{gi}\sqrt{c_i})^2/C_f$, we obtain a simpler form, because $V_{gi}^2 c_i / V_{g\min} = (V_{gi}\sqrt{c_i}/\sum V_{gi}\sqrt{c_i})^2$. Thus the jointly determined minimal loss function becomes

$$\begin{aligned} 1+L(m_i^{**}) &= \left[\sum_i \sqrt{\left\{\sum_g I_g \left(V_{gi}\sqrt{c_i} \Big/ \sum_i V_{gi}\sqrt{c_i}\right)^2\right\}}\right]^2 \\ &= \left\{\sum_i \sqrt{\left(\sum_g I_g U_{gi}^2\right)}\right\}^2. \end{aligned} \tag{6.7}$$

The minimal and optimal values may be unobtainable, due chiefly to the constraints $m_i^* \leqslant M_i$ (Section 5). In that case the above loss function overestimates the losses incurred over obtainable values of $V_{g\min}$. Note also that using these leads to $Z_i\sqrt{c_i} = \sqrt{(\sum_g I_g U_{gi}^2)}\sqrt{C_f}$, hence to

$$\text{optimal } (m_i^{**}) = \frac{\sqrt{(\sum I_g U_{gi}^2)}}{\sum \sqrt{(\sum I_g U_{gi}^2)}} \frac{C_f}{c_i}. \tag{6.8}$$

These can be seen applied in Section 7.6 to the important and frequent conflict between allocations for weighted totals and for comparisons of domains. Two examples are shown in Table 7.5. Note in the last column of Table 5(B) how encouragingly insensitive are the values of (6.7) for moderate differences in the assignments of I_g.

The weighted mean of relative quadratic losses (6.3) is a modified version of a function proposed by Dalenius (1957, Chapter 9). Another version (Yates, 1960; Cochran, 1963) uses $\sum_g I_g' \sum_i V_{gi}^2/m_i$ the weighted average of variances. Our (6.3) can be easily adapted by using $T_i^2 = \sum_g I_g' V_{gi}^2$ instead of Z_i^2; in this formulation the weights $I_g' = I_g/V_{g\min}$ include the minimal variances. This may appear simpler, but it is less explicit.

The optimal allocation of $m_i^* \propto Z_i/\sqrt{c_i}$ can also be obtained with Lagrange multipliers applied to the function

$$F(m_i) = \sum_g I_g \sum_i V_{gi}^2 \Big/ m_i V_{g\min} + \lambda \sum c_i m_i. \tag{6.9}$$

With Lagrange multipliers we also investigated two other loss functions: the product, $\prod(1+L_g)$, and the sum of the relative precisions, $\{\sum(1+L_g)^{-1}\}^{-1}$. But the results seem less crucial than good choices for the weights I_g of relative importance.

Our methods here aim to minimize the first term of $V+V_0$ for fixed C_f. In situations where V_0 is considerable, the actual loss should be modified to $L' = L/(1+V_0/V_{\min})$, as noted in Section 3. Furthermore, I consider fixing C_f more practical than trying to fix values for a set of V_g and then to minimize C_f. This problem seems to have been solved with "convex programming" on several separate occasions (Srikantan, 1963, and Hartley, 1965, for example), but I do not find this approach useful.

7. Several Applications

7.1. *Stratified Element Sampling*

In this common application the variance and cost functions can be written, with $\Sigma W_i = 1$, as $V+V_0 = \Sigma(W_i S_i)^2/m_i - \Sigma(W_i S_i)^2/M_i$ and $C+C_0 = \Sigma c_i m_i + C_0$. Here $V_i = W_i S_i$ and $U_i = W_i S_i \sqrt{c_i}/\Sigma W_i S_i \sqrt{c_i}$. Often the c_i and sometimes the S_i^2 are treated as equal among strata; then the $U_i = W_i S_i/\Sigma W_i S_i$ or $U_i = W_i$. To these parameters the earlier results on optima and proxima can be applied.

For a fixed $m = \Sigma m_i = \Sigma m_i^*$, disregarding differences in the c_i, the loss function (2.7) with sample weights becomes (also shown in Cochran, 1963, 5A.1), with $d_i = (m_i^* - m_i)/m_i$:

$$L = \frac{V_{\text{actual}}}{V_{\min}} - 1 = \Sigma \frac{m_i}{m} \frac{(m_i^* - m_i)^2}{m_i^2} = \frac{1}{m} \Sigma \frac{(m_i^* - m_i)^2}{m_i} = (1/m) \Sigma m_i d_i^2. \tag{7.1}$$

7.2. *Two Components; Subsampling Non-responses*

For only two components (e.g. two strata or two stages) we can obtain several convenient applications. Optimal numbers of units stand in the ratio

$$\frac{m_2^*}{m_1^*} = \frac{V_2}{V_1} \sqrt{\left(\frac{c_1}{c_2}\right)}. \tag{7.2}$$

Optimal allocations result in

$$\text{either} \quad V_{\min}^2 = (V_1 \sqrt{c_1} + V_2 \sqrt{c_2})^2 / C_f \quad \text{or} \quad C_{\min} = (V_1 \sqrt{c_1} + V_2 \sqrt{c_2})^2 / V_f. \tag{7.3}$$

Now let $U = V_1 \sqrt{c_1}/(V_1 \sqrt{c_1} + V_2 \sqrt{c_2})$ and $(1-U)$ denote the two population weights; and let $K = k_2/k_1$ denote the *relative* departure from optimal numbers, so that $m_1/m_2 = K m_1^*/m_2^*$. Choosing the order of components is immaterial for expressing the loss L (from (2.3)), due to departure from optimal, as

$$1+L = (UK+1-U)(U/K+1-U) = 1+U(1-U)\{(K-1)^2/K\}. \tag{7.4}$$

For computing the weights we may use any of the relationships

$$\frac{U}{1-U} = \frac{V_1 \sqrt{c_1}}{V_2 \sqrt{c_2}} = \frac{m_1^* c_1}{m_2^* c_2} = K \frac{m_1 c_1}{m_2 c_2} = r; \quad \text{then} \quad U = r/(1+r). \tag{7.5}$$

The loss is greatest with $U = 0{\cdot}5$ for fixed K. It is modest for $0{\cdot}5 < k < 2$, but rises sharply for extreme values of K (see Table 1). For example, for a stratified sample we may often assume $S_2/\sqrt{c_2} = S_1/\sqrt{c_1}$, and optimal allocation would be $m_2^*/m_1^* = (1-W_1)/W_1$. But if we actually have sample numbers in the ratio $m_2/m_1 = 4(1-W_1)/W_1$, the departure $K = 4$ results in the loss $(9/4) W_1(1-W_1)$. See column $K = 4$ of Table 1; also example 2(a) in Section 2.

A ready application is to subsampling of non-responses. Suppose that m questionnaires are mailed out for a total cost of $c_0 m$; that Rm respond to several mailings for a further cost of $c_r Rm$; and that a fraction $1/k$ of the $(1-R)m$ non-responses are interviewed for a cost of $c_n(1-R)m/k$.

Thus,

$$\text{cost}(\bar{y}) = (c_0/R + c_r) Rm + c_n(1-R)m/k. \tag{7.6}$$

Assume either that the m were selected with *srs* or that design effects are included. Denote with S_r^2 and S_n^2 the element variances of response and non-response. The variance may be written, without the constant factor for finite population, approximately as

$$\text{var}(\bar{y}) = \frac{R^2 S_r^2}{Rm} + \frac{(1-R)^2 S_n^2}{(1-R)m/k}. \tag{7.7}$$

From (7.2) we get

$$\text{optimal } k^* = \frac{S_r}{S_n}\sqrt{\left(\frac{c_n}{c_0/R + c_r}\right)}. \tag{7.8}$$

7.3. *Two-stage and Multistage Selection*

In two-stage random selection without replacement from equal clusters we have

$$\text{var}(\bar{y}) = \left(1-\frac{a}{A}\right)\frac{S_a^2}{a} + \left(1-\frac{b}{B}\right)\frac{S_b^2}{ab} = \frac{(S_a^2 - S_b^2/B)}{a} + \frac{S_b^2}{ab} - \frac{S_a^2}{A}, \tag{7.9}$$

where the last term denotes the constant V_0; and $\text{cost}(\bar{y}) = c_a a + c_b ab + C_0$. Using (7.2) the ratio $m_2^*/m_1^* = ab/a$ yields

$$\text{optimal } b^* = \frac{S_b}{S_u}\sqrt{\left(\frac{c_a}{c_b}\right)}, \tag{7.10}$$

where $S_u^2 = S_a^2 - S_b^2/B$ yields a shorter form that is meaningful: the cluster variance additional to average random variance of its B elements. $S_u^2 \leqslant 0$ would cause a mathematical dilemma because we have assumed positive components, but in practice it leads to taking all B elements, selecting complete clusters. This design $b = B$ also serves when (7.10) yields $b^* > B$; and it may be a practical solution even when b^* is a large fraction of B. In some situations one could search for a new definition of larger clusters. These departures from optimal b^* interfere with attaining the optimal values of $V_{\min}$ and $C_{\min}$ (see Section 5).

Extensions to three (or more) stages are not difficult:

$$\text{var}(\bar{y}) = \frac{(S_a^2 - S_b^2/B)}{a} + \frac{(S_b^2 - S_c^2/C)}{ab} + \frac{S_c^2}{abc} - \frac{S_a^2}{A}$$

$$= \frac{S_u^2}{a} + \frac{S_w^2}{ab} + \frac{S_c^2}{abc} - \frac{S_a^2}{A}. \tag{7.11}$$

7.4. *Two-phase Sampling*

Suppose a large sample of size n_L is selected (with *srs*) from N in the first phase, and a sample of size n in the second phase. The cost function may be written as

$$\text{cost}(\bar{y}) = C_0 + cn + c_L n_L, \tag{7.12}$$

where c_L and n_L denote unit cost and numbers for the large first phase sample, and c and n for the smaller second phase.

When used for proportionate stratification, $n = \sum n_h$, and the variance is to a good approximation

$$\text{var}(\bar{y}) = \sum W_h^2 \frac{S_h^2}{n_h} + \frac{1}{n_L}\sum W_h(\bar{Y}_h - \bar{Y})^2 - \frac{\sum W_h^2 S_h^2}{N_h} - \frac{\sum W_h(\bar{Y}_h - \bar{Y})^2}{N}. \tag{7.13}$$

When fixed $n = \sum n_h$ is allocated optimally according to $n_h^*/n = W_h S_h / \sum W_h S_h$, $V_{\min}$ becomes

$$\text{var}(\bar{y}) - V_0 = \frac{(\sum W_h S_h)^2}{n} + \frac{1}{n_L}\sum W_h(\bar{Y}_h - \bar{Y})^2. \tag{7.14}$$

Allocation of (7.2) and (7.3) between the two cost components of $C = cn + c_L n_L$ yields

$$\text{optimal}\,\frac{n_L}{n} = \sqrt{\left\{\frac{\sum W_h(\bar{Y}_h - \bar{Y})^2}{(\sum W_h S_h)^2}\,\frac{c}{c_L}\right\}} \quad \text{and} \quad V_{\min} = [\sum W_h S_h \sqrt{c} + \sqrt{\{\sum W_h (Y_h - Y)^2 c_h\}}]^2/C_f. \tag{7.15}$$

If the cost factors c_h vary among the H strata, the optimal n_h^* should be made proportional to $W_h S_h/\sqrt{c_h}$. We may solve directly for the $(H+1)$ unknowns (n_h and n_L) by applying (5.2) and (5.3) to $C = \sum n_h c_h + c_L n_L$, and to (7.13). The first term becomes $\sum W_h S_h \sqrt{c_h}$ in $V_{\min}$.

In two-phase sampling for regression estimation, the variance may be expressed to a good approximation as

$$V^2_{(\bar{y})} = \frac{S^2(1 - R^2_{yx})}{n} + \frac{R^2_{yx} S^2}{n_L}. \tag{7.16}$$

Allocation of the two cost components yields

$$\text{optimal}\,\frac{n_L}{n} = \frac{R_{yx}}{\sqrt{1 - R^2_{yx}}}\,\frac{\sqrt{c}}{\sqrt{c_L}} \quad \text{and} \quad V_{\min} = S^2\{\sqrt{(1 - R^2_{yx})}\sqrt{c} + R_{yx}\sqrt{c_L}\}^2/C_f. \tag{7.17}$$

7.5. *Weights in Estimation*

Our emphasis has been in the allocation of units m_i in selection, but the method can also be applied to the allocation of weights w_i in linear estimation. Let y_i $(i = 1, 2, \ldots, n)$ be independent estimates of $\bar{Y}$, with $\text{var}(y_i) = \sigma_i^2$ (with $0 < \sigma_i < \infty$). Suppose one estimates $\bar{Y}$ with a weighted mean $\bar{y} = \sum w_i y_i$, with the constraint $\sum w_i = 1$. How to choose the w_i to minimize the variance of $\sum w_i y_i$? We denote $\text{var}(\sum w_i y_i) = \sum w_i^2 \sigma_i^2 = V$, and corresponding to VC we form the product

$$VW = (\sum w_i^2 \sigma_i^2)(\sum w_i) = (\sum w_i \sigma_i \sqrt{w_i})^2 + \sum_{i<j} (w_i \sigma_i \sqrt{w_j} - w_j \sigma_j \sqrt{w_i})^2. \tag{7.18}$$

The optimal values w_i^* are reached when $\sqrt{w_i^*}\,\sigma_i = F$, thus

$$w_i^* = F^2/\sigma_i^2 \quad \text{and} \quad \sum w_i^* = F^2 \sum 1/\sigma_i^2 = 1.$$

Then

$$V_{\min} = \sum w_i^* F^2 = F^2 = (\sum 1/\sigma_i^2)^{-1} \quad \text{and} \quad w_i^* = \frac{1/\sigma_i^2}{\sum 1/\sigma_i^2}. \tag{7.19}$$

The relative loss L may be found from

$$1 + L = V/V_{\min} = (\sum 1/\sigma_i^2)(\sum w_i^2 \sigma_i^2) = \sum \frac{w_i^2}{w_i^*}. \tag{7.20}$$

Thus

$$L = \sum \frac{(w_i - w_i^*)^2}{w_i^*} = \sum w_i^* \left(\frac{w_i - w_i^*}{w_i^*}\right)^2 = \sum w_i^* (k_i' - 1)^2, \tag{7.21}$$

with $\sum w_i^* k_i' = 1$, since $\sum w_i^* = \sum w_i = 1$. The loss appears as the relvariance of the relative departures k_i' (from mean of 1) with weights w_i^*—see applications to standardization by Kalton (1968).

For equal variances, $\sigma_i^2 = \sigma^2$, the optimal values are also equal, $w_i^* = 1/n$, and $V_{\min} = \sigma^2/n$, and the loss function becomes $1 + L = (1 + C_k^2)$, where C_k^2 is the relvariance among the relative departures. When there are only a few values of k_i with relative frequencies W_i, use $(1 + L) = (\sum W_i k_i^2)/(\sum W_i k_i)^2$.

The loss due to unequal weights, C_k^2, represents the discrepancy between them. Therefore, when random replication is used to obtain integral weights (machine cards) from fractions $k+W$, the weighting should be confined to successive integers; with k applied to $(1-W)$ and $(k+1)$ to W of the elements, we have $(1-W)k+W(k+1)=k+W$ cards. The variance is increased by

$$1+L=\frac{(1-W)k^2+W(k+1)^2}{(k+W)^2}=1+\frac{W(1-W)}{(k+W)^2}. \tag{7.22}$$

Duplicating for non-response is common, with $k=1$. The loss then is

$$L=W(1-W)/(1+W)^2;$$

its maximum is 0·125, when $W=\frac{1}{3}$. The case of $k=0$ would refer to eliminating $(1-W)$, leading to a loss of $L=(1-W)/W$. Note that eliminating a small fraction $(1-W)$ surprisingly appears only a little worse than duplicating a similarly small fraction W. For example, eliminating 0·05 results in $L=0{\cdot}056$; duplicating 0·05 results in $L=0{\cdot}046$. Hence to equalize several groups with k_i that differ only slightly, instead of duplicating up to the highest response, one may reduce L by eliminating from the groups with the highest response.

7.6. *Allocation Conflict between Totals and Independent Domains*

Serious conflict often exists between reducing the variance for the combined mean $\sum W_i\bar{y}_i$, and equal precision desired for the means $\bar{y}_i$ of H independent domains that differ greatly in relative sizes $W_i(\sum W_i=1)$. The domains may be the regions or provinces of a country, etc. This common example of multipurpose allocation deserves special attention.

The *combined mean variance* $V_c=\sum W_i^2S_i^2/m_i$ is minimal when the optimal $m_{ci}^*\propto W_iS_i/\sqrt{c_i}$. However, $m_{di}^*\propto S_i/\sqrt{c_i}$ are optimal for obtaining equal precision for each of the H domain means; also to obtain equal precision for the $H(H-1)/2$ possible comparisons of domain means. Thus we can denote an *average domain variance* $V_d=(\sum S_i^2/m_i)/H^2$ for the variance of $\sum\bar{y}_i/H$. The conflict between the purposes is represented in the above two optimal values for m_i^* by the presence of the weights W_i for the combined mean, and their absence for the domain means. Thus the loss function (2.3) for the combined mean, due to allocations $m_i\propto S_i/\sqrt{c_i}$, has the departures $k_{ci}=m_i^*/m_i=W_i$, and the weights $U_{ci}\propto W_iS_i\sqrt{c_i}$. The loss function for the average domain means, due to allocations $m_i\propto W_iS_i\sqrt{c_i}$, has the departures $k_{di}=1/W_i$ and the weights $U_{di}\propto S_i\sqrt{c_i}$.

To see clearly the effects of variation in the domain sizes W_i, we make some simplifying assumptions that are often approximated in practical situations. Assume that the S_i^2 incorporate the effects of complex designs, and that they are constant across domains, as are the c_i. Further, suppose that $m_i^*\leqslant M_i$ in all domains. We shall also neglect effects of the constants V_0 and C_0 on the loss functions.

Under these conditions we may omit, for brevity, the constants S^2 and c from the formulae, and we allocate the total sample size $m=\sum m_i$ among the domains. For $\sum W_i\bar{y}_i$ the optimal $m_i^*=mW_i$, with departures $k_{ci}=m_i^*/m_i=mW_i/m_i$ and weights $U_{ci}=W_i$, the loss function $1+L_c=m\sum W_i^2/m_i$ is minimal at $mV_{c\min}=1$. For $\sum\bar{y}_i/H$ the optimal $m_{di}^*=m/H$, weights $U_{di}=1/H$, with departures $k_{di}=m/Hm_i$ and the loss function $1+L_d=mH^{-2}\sum 1/m_i$ is minimal at $mV_{d\min}=1$ also.

In Table 4 for loss functions $(1+L_c)$ the minimal value 1 appears with $m_i\propto W_i$ for $\sum W_i\bar{y}_i$, and with $m_i\propto 1/H$ for $\sum\bar{y}_i/H$. The other allocations produce relative losses $(L>0)$ that increase with diversity among the relative sizes W_i; and C_w^2 denotes their relative variance, $H^2\operatorname{var}(W_i)$.

Jointly for the two purposes, we can find optimal allocation and the loss function with (6.3). For any allocation m_i, the joint function is

$$1+L_j(m_i)=I_cm\sum W_i^2/m_i+I_dmH^{-2}\sum 1/m_i=m\sum(I_cW_i^2+I_dH^{-2})/m_i=m\sum_i t_i^2/m_i, \tag{7.23}$$

where

$$t_i = \sqrt{(I_c W_i^2 + I_d H^{-2})} = \sqrt{(I_c D_i^2 + I_d)}/H = \sqrt{(I_c N_i^2 + I_d \bar{N}^2)}/N.$$

Here $0 < I_c < 1$ is the relative importance for the combined mean variance and $I_d = 1 - I_c$ for the mean domain variance. We may find it convenient to use $D_i = HW_i$ with mean $\bar{D} = 1$, or $N_i = NW_i$ when these denote domain sizes and $N = \sum N_i = H\bar{N}$.

We can find the joint optimal allocations $m_i^{**} = mt_i/\sum t_i$ simply with (5.3), but also as an illustration of (6.8).

TABLE 4

Conflict of Combined Mean $(\sum W_i \bar{y}_i)$ *and Average Domain Mean* $(\sum \bar{y}_i/H)$
$(S_i^2$ *and* c_i *are assumed constant and omitted)*
Loss function $(1+L)$ *for the combined mean, for the average domain mean, and for a weighted joint function*
Note $t_i = \sqrt{(I_c W_i^2 + I_d H^{-2})} = \sqrt{(I_c D_i^2 + I_d)}/H$

	Loss functions $(1+L)$ *for*		
$(1+L) = mV^2$	$\sum W_i \bar{y}_i$ $m \sum W_i^2/m_i$	$\sum \bar{y}_i/H$ $mH^{-2} \sum 1/m_i$	$I_c \sum W_i \bar{y}_i + I_d \sum \bar{y}_i/H$ $m \sum t_i^2/m_i$
Allocation of m_i			
mW_i	1	$H^{-2} \sum 1/W_i$	$I_c + I_d H^{-2} \sum 1/W_i$
m/H	$H \sum W_i^2 = 1 + C_w^2$	1	$I_c H \sum W_i^2 + I_d H^{-1}$
$mt_i/\sum t_i$	$(\sum W_i^2/t_i)(\sum t_i)$	$H^{-2}(\sum 1/t_i)(\sum t_i)$	$(\sum t_i)^2$

TABLE 5

Loss functions $(1+L)$ *for two populations*

	(A) $(1+L)$ *for* $W_1/W_2 = 4$			(B) $(1+L)$ *for* 133 *countries*: 0·2 *to* 100 *mm*			
						Joint with weights	
Allocations m_i	$\sum W_i \bar{y}_i$	$\sum \bar{y}_i/2$	*Joint*	$\sum W_i \bar{y}_i$	$\sum \bar{y}_i/133$	1 : 1	$I_c/I_d : 1$
mW_i	1	1·56	1·28	1	6·86	3·93	
m/H	1·36	1	1·18	3·34	1	2·17	
$\propto \sqrt{W_i}$	1·08	1·125	1·102	1·35	1·54	1·44	
$\propto \sqrt{(W_i^2 + H^{-2})}$	1·116	1·080	1·098	1·31	1·28	1·295	
$\propto \sqrt{(0{\cdot}5W_i^2 + H^{-2})}$				1·47	1·17	(1·32)	1·27
$\propto \sqrt{(2W_i^2 + H^{-2})}$				1·20	1·44	(1·32)	1·28
$\propto \sqrt{(4W_i^2 + H^{-2})}$				1·12	1·66	(1·39)	1·23

In (A) there are two strata and domains ($W_1 = 0{\cdot}8$ and $W_2 = 0{\cdot}2$); note that the allocation $m_i = \sqrt{W_i}$ does almost as well for the joint loss as the optimal.

In (B) we have the populations of 133 countries, ranging in size from 0·2 to over 100 millions, a range of 500 in relative sizes. From this problem of allocation (for the World Fertility Survey) we omitted, for practical reasons, the four largest countries and a few under 0·2 millions. Their inclusion would raise the variance of relative sizes, W_i, from 2·5 to 12, and would make the results more dramatic. Note that the $\sqrt{W_i}$ allocation reduces losses quite well. Some compromise is better than none. But the optimal allocation, $\sqrt{(W_i^2 + H^{-2})}$, is considerably better. Different values of I_c/I_d (= 1/2, 2/1 and 4/1) increase slightly the variance of the joint loss function with (1 : 1) weights; but they remain steady for joint loss functions with their own weights $I_c/I_d : 1$.

The multipurpose allocation m_i^{**} can also be shown (5.2) to produce the multipurpose minimal variance

$$V_{\min} = (\sum t_i)^2/m. \tag{7.24}$$

When we use the multipurpose optimal $m_i^{**} \propto t_i$ we can determine the loss functions $(1+L)$ incurred for the variances of $\sum W_i \bar{y}_i$ and $\sum \bar{y}_h/H$; we use (6.1) or (6.2) with $k_{ci} \propto V_{ci}/V'_i \propto W_i/t_i$ and $k_{di} \propto V_{di}/V'_i \propto 1/Ht_i$ respectively. These $(1+L)$ are shown on the bottom row of Table 4. The last column shows the effects of the three different allocations on the joint multipurpose loss function $1+L_j(m_i)$.

Two numerical problems illustrate the method in Table 5. In (A), for two domains having sizes $W_1/W_2 = 4:1$, are shown the loss functions for three purposes—total, domain and joint —under diverse allocations. In (B) the method is applied to the 133 countries of the world, omitting the four largest, over 200 millions, and a few smallest, under 0·2 millions. Including them would be more dramatic but less realistic.

ACKNOWLEDGEMENTS

The work was supported by Grants GS-777 and GS-3191X from the National Science Foundation. I received help from W. G. Cochran, W. H. DuMouchel and from colleagues in 1969 and 1972 in the Statistics Department of the London School of Economics, especially D. R. Brillinger and G. Kalton. The referee was very helpful.

REFERENCES

COCHRAN, W. G. (1963). *Sampling Techniques*, 2nd ed. New York: Wiley.

DALENIUS, T. (1957). *Sampling in Sweden.* Stockholm: Almquist and Wicksell.

EVANS, W. D. (1951). On stratification and optimum allocation. *J. Amer. Statist. Ass.*, **30**, 219–229.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory*, Vols 1 and 2. New York: Wiley.

HARTLEY, H. O. (1965). Multiple purpose optimum allocation in stratified sampling. *Proc. Soc. Statist. Sect. Amer. Statist. Ass.*, 258–261.

HUDDLESTON, H. F., CLAYPOOL, P. L. and HOCKING, R. R. (1970). Optimal sample allocation to strata using convex programming. *Appl. Statist.*, **19**, 273–278.

KALTON, G. (1968). Standardization: a technique to control for extraneous variables. *Appl. Statist.*, **17**, 118–136.

KISH, L. (1961). Efficient allocation of a multipurpose sample. *Econometrica*, **29**, 363–385.

—— (1965). *Survey Sampling.* (Especially Sections 8.5 and 11.7.) New York: Wiley.

—— (1969). Design and estimation for subclasses, comparisons and analytical statistics. Chapter 21 in *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, eds). New York: Wiley.

KISH, L. and FRANKEL, M. R. (1974). Inference from complex samples (with Discussion). *J. R. Statist. Soc.* B, **36**, 1–37.

KOKAN, A. R. and KHAN, S. (1967). Optimum allocation in multivariate surveys; an analytical solution. *J. R. Statist. Soc.* B, 115–125.

SRIKANTAN, K. S. (1963). A problem in optimum allocation. *Operat. Res.*, **11**, 265–273.

STUART, A. (1954). A simple presentation of optimum sample results. *J. R. Statist. Soc.* B, **16**, 239–241.

SUKHATME, P. V. and SUKHATME, B. V. (1970). *Sampling Theory of Surveys with Applications*, 2nd ed. Ames: Iowa State University Press.

TUKEY, J. W. (1948). Approximate weights. *J. Amer. Math. Soc.*, **19**, 91–92.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, London: Griffin.