# OPTIMAL AND PROXIMAL MULTIPURPOSE ALLOCATION

Leslie Kish, The University of Michigan

## 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 $\bar{y}$:

$$\mathrm{Var}(\bar{y}) = V + V_o = \Sigma V_i^2/m_i + V_o \qquad (1.1)$$

and

$$\mathrm{Cost}(\bar{y}) = C + C_o = \Sigma c_i m_i + C_o \,. \qquad (1.2)$$

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 $\bar{y} = \Sigma W_i \bar{y}_i$ is:

$$\mathrm{Var}(\bar{y}) = \Sigma (W_i S_i)^2/m_i - \Sigma (W_i S_i)^2/M_i ,$$

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

$$\mathrm{Var}(\bar{y}) = (1 - a/A)S_a^2/a + (1 - b/B)S_b^2/ab$$

$$= S_u^2/a + S_b^2/ab - S_a^2/A.$$

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 $-S_a^2/A = V_o$ does not depend on the $m_i$. The cost is $c_a a + c_b ab + C_o$.

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.

(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 $V_i \geq 0$ and $\sqrt{c_i} \geq 0$ (hence $V_i^2$ and $c_i$) for allocating the $m_i$. For nontrivality 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 7C.

(6) The constants $V_o$ and $C_o$ do not affect optimal allocations of the $m_i$; their effects on losses in proximal allocation are shown in Section 3. $C_o$ is nonnegative, but $V_o$ is often negative, as above.

(7) For practical values of $m_i$ we want positive integers. Also $0 < m_i \leq M_i$, where $M_i$ denotes the number of units in the population for the component; and $m_i \geq 2$ for computing variance components. Frequently allocation formulas 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 - 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 worthwhile. 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 $C/c = m = \Sigma m_i$. Then the $\sqrt{c_i}$ should be omitted from the allocation formulas. Instead of $C_o$ use $C_o/c$, where c is a common (average) unit cost.

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

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

$$VC = (\Sigma V_i^2/m_i)\,(\Sigma c_i m_i)\,, \qquad (1.3)$$

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

$$\mathrm{Var}(\bar{y}) \times \mathrm{Cost}(\bar{y}) = (V + V_o)\,(C + C_o),$$

* Support from the National Science Foundation grant 3191X and suggestions from several colleagues were gratefully received. Curtailed version of contribution to JRSS(A) 1975, which includes the references.

because in $(V + V_o)C_f$ or in $V_f(C + C_o)$ 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.

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 7F, and fulfill our second aim.

Third, 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 $(\Sigma V_i\sqrt{c_i})^2$; 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" $U_i = V_i\sqrt{c_i}/\Sigma V_i\sqrt{c_i}$.

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

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

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

where $U_i = V_i\sqrt{c_i}/\Sigma V_i\sqrt{c_i}$ 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 $\Sigma U_i = 1$; and we may also use any convenient $U_i' \propto U_i$, if we divide by $\Sigma U_i'$. 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 4.1 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 $\Sigma W_i\bar{y}_i$ for two strata where $W_1 = 0.2$, $W_2 = 0.8$, $S_1^2 = S_2^2 = S^2$, 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 $1 + L = (0.2\cdot 1 + 0.8\cdot 1/4)(0.2\cdot 1 + 0.8\cdot 4) = 1.360$.

(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 $U_i = W_i S_i\sqrt{c_i}/\Sigma W_i S_i\sqrt{C_i}$; hence $U_1 = U_2 = 0.5$. Therefore the relative loss L from equal sample sizes now would be given by $1 + L =$

$$(0.5\cdot 1 + 0.5\cdot 1/4)(0.5\cdot 1 + 0.5\cdot 4) = 1.5625.$$

(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 $U_1 = U_2 = 0.5$. Optimal allocation is at $m_1 = m_2$. Departure from this in the ratio 1:4 to satisfy (a) would result in $1 + L =$

$$(0.5\cdot 1 + 0.5\cdot 4)(0.5\cdot 1 + 0.5\cdot 1/4) = 1.5625.$$

Note that these answers can also be found in Table 4.1 in column K = 4 for relative departures. The size 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 7F, and in Table 7.2A.

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:

$$1 + L = (\Sigma u_i k_i^2)(\Sigma u_i) \quad (2.4)$$

$$= 1 + (\Sigma u_i k_i^2/\Sigma u_i - \bar{k}^2)/\bar{k}^2 = 1 + C_k^2 \quad (2.5)$$

$$= 1 + \Sigma u_i (k_i/\bar{k} - 1)^2/\Sigma u_i \quad (2.6)$$

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

$C_k^2$ is the relative variance of the $k_i$ with sample weights $u_i$ around their mean $\bar{k} = \Sigma u_i k_i/\Sigma u_i = 1/\Sigma u_i$. 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 $U_i = V_i\sqrt{c_i}/\Sigma V_i\sqrt{c_i} \propto c_i m_i^* \propto c_i m_i k_i$.

## 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 practicing statisticians are given formal and practical expressions (2.4 - 2.7) in terms of the relative loss (L) compared to optimal allocation.

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 formulas 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 4.1 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 formulas and tables take account of the factors V and C in the minimized function VC, but they neglect the constant $V_o$ or $C_o$ 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 $V_o$ and $C_o$. If $V_{min}$ is the optimal V for fixed $C_f^o$ then the ratio of the attained proximal variance to the optimal variance is

$$\frac{(1+L)V_{min} + V_o}{V_{min} + V_o} = 1 + L/(1 + V_o/V_{min}) = 1 + L' . \quad (3.1)$$

Thus the adjusted actual relative loss L' differs from that indicated by L; since $V_o$ 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 $C_o$ in

$$L' = L/(1 + C_o/C_{min}) \quad (3.2)$$

## 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

$$L = U(1-U)(K-1)^2/K. \quad (4.1)$$

The dichotomous models represent maximal losses for ranges of departure fixed at 1 to K. Thus losses for large values of K are much greater in the top three rows of Table 4.1 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 = 1.3 $L_d = .017$ and $L_c = .006$, and for K = 1.5 $L_d = .04$ and $L_c = .014$. (Note that for $L_d$ the $k_i$ take only two values 1 and K = 1.3 or 1.5). 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.

Table 4.1 Relative Losses (L) for 6 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 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dichotomous U(1-U) | | | | | | | | | | | | | |
| (0.5)(0.5) | | .017 | .042 | .125 | .333 | .562 | .800 | 2.025 | 4.512 | 12.005 | 24.50 | 124.5 | 249.5 |
| (0.2)(0.8) | | .011 | .027 | .080 | .213 | .360 | .512 | 1.296 | 2.888 | 7.683 | 15.68 | 79.7 | 159.7 |
| (0.1)(0.9) | | .006 | .015 | .045 | .120 | .202 | .288 | .729 | 1.624 | 4.322 | 8.82 | 44.8 | 89.8 |
| Rectangular | $L_d$ | .017* | .042* | .125* | .222 | .302 | .370 | .611 | .889 | 1.295 | 1.620 | 2.403 | 2.746 |
| $U_i \propto 1/K$ | $L_c$ | .006 | .014 | .040 | .099 | .155 | .207 | .407 | .656 | 1.036 | 1.349 | 2.120 | 2.461 |
| Linear Decrease | $L_d$ | .017* | .040* | .111* | .203 | .283 | .353 | .616 | .940 | 1.437 | 1.917 | 2.879 | 3.333 |
| $U_i \propto K+1-k_i$ | $L_c$ | .006 | .014 | .040 | .097 | .153 | .205 | .409 | .680 | 1.127 | 1.514 | 2.507 | 2.956 |
| Hyperbolic Decr. | $L_d$ | .017* | .040* | .111* | .215 | .312 | .404 | .807 | 1.466 | 3.014 | 5.076 | 16.802 | 28.342 |
| $U_i \propto 1/k_i$ | $L_c$ | .006 | .014 | .041 | .103 | .171 | .235 | .528 | 1.011 | 2.138 | 3.621 | 11.998 | 19.915 |
| Quadratic Decr. | $L_d$ | .016* | .036* | .080* | .150 | .211 | .264 | .460 | .696 | 1.048 | 1.333 | 2.026 | 2.331 |
| $U_i \propto 1/k_i^2$ | $L_c$ | .006 | .014 | .040 | .099 | .155 | .207 | .407 | .656 | 1.036 | 1.349 | 2.120 | 2.461 |
| Linear Increase | $L_d$ | .017* | .040* | .111* | .167 | .200 | .222 | .273 | .302 | .320 | .327 | 0.330 | 0.333 |
| $U_i \propto k_i$ | $L_c$ | .006 | .013 | .037 | .083 | .120 | .148 | .223 | .273 | .308 | .320 | 0.331 | 0.332 |

Dichotomous $1 + L = 1 + U(1-U)(K-1)^2/K$

Discrete $1 + L_d = (\Sigma U_i k_i)(\Sigma U_i/k_i)$, with $k_i = i = 1,2,3,\ldots K$

Continuous $1 + L_c = \int Uk.dk.\int (U/k)dk$, with $1 \le k \le K$.

Only 2 Values, 1 and K, were used for $L_d$ for K = 1.3, 1.5 and 2

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 $L_d = 0.370$; more than 5 values evenly spaced within the same range to 1 to 5 would have a loss between that value and the continuous loss $L_c = .207$. On the other hand for only three values of $k_i$ from 1 to 5 and $U_i = 1/3$, the loss (actually 0.533) is above 0.370, but below the dichotomous value of 0.800 in Table 4.1.

When <u>sample weights</u> $u_i = U_i/k_i$ 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). Formulas and tables can be constructed for such relvariances, if we begin with the means M and variances $\sigma^2$ of convenient distributions from 0 to 1 [Kish, 1965, p.262]. To obtain the relvariances $C_k^2$, those variances are multiplied by the $(\text{range})^2 = (K-1)^2$ and divided by the new $(\text{mean})^2 = [M(K-1)+1]^2$; thus $C_k^2 = \sigma^2(K-1)^2/[M(K-1)+1]^2$.

## 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 nonnegative values. Then

$$(\Sigma x_i^2)(\Sigma y_i^2) = \Sigma x_i^2 y_i^2 + \sum_{i \neq j} x_i^2 y_j^2$$

$$= \Sigma x_i^2 y_i^2 + \sum_{i \neq j} x_i y_i x_j y_j + \sum_{i \neq j} x_i^2 y_j^2 - \sum_{i \neq j} x_i y_i x_j y_j$$

$$= (\Sigma x_i y_i)^2 + \sum_{i<j} (x_i y_j - x_j y_i)^2. \quad (5.1)$$

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

$$1 + L = (\Sigma U_i k_i)(\Sigma U_i k_i) = \Sigma y_i^2 \Sigma x_i^2$$

$$= 1 + \sum_{i<j} \frac{U_i U_j}{k_i k_j}(k_i - k_j)^2, \quad (5.1')$$

with $(\Sigma x_i y_i)^2 = (\Sigma U_i)^2 = 1$. 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 $\ge 0$). The minimal value of

$$VC = (\Sigma V_i^2/m_i)(\Sigma c_i m_i) \ge (\Sigma V_i \sqrt{c_i})^2, \quad (5.2)$$

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

Then
$$m_i^* = FV_i/\sqrt{c_i} \quad (5.3)$$
are the optimal values of the $m_i^*$ that obtain the

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

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

## 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 $\Sigma V_i^2/m_i$ for variances can serve well the many purposes of a sample survey; for the gth purpose the variance will be denoted by $\Sigma_i V_{gi}^2/m_i$.

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 7F.

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 $U_i = V_i\sqrt{c_i}/\Sigma V_i\sqrt{c_i}$ 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

$$1 + L(m_i') = (\Sigma V_i^2\sqrt{c_i}/V_i')(\Sigma V_i'\sqrt{c_i})/(\Sigma V_i\sqrt{c_i})^2 \quad (6.1)$$

$$= \sum \left(\frac{V_i\sqrt{c_i}}{\Sigma V_i\sqrt{c_i}}\right)^2 \Bigg/ \left(\frac{V_i'\sqrt{c_i}}{\Sigma V_i'\sqrt{c_i}}\right).$$

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

$$1 + L(m_i') = \Sigma(V_i/\Sigma V_i)^2/(V_i'/\Sigma V_i'). \quad (6.1')$$

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

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

Now consider a loss function for several variates indexed with $g(= 1, 2, 3 \ldots)$. The loss function, for a fixed cost $C_f = \Sigma c_i m_i$, may be expressed for each as

$$1 + L_g = (\Sigma V_{gi}^2/m_i)/V_{gmin},$$

where the denominator denotes the minimal variance attainable and computed for the gth variate. Assign the weights $I_g$ $(\Sigma I_g = 1)$ to denote the relative importance of the lost precision on 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

$$1 + L(m_i) = \Sigma_g I_g(1 + L_g) = 1 + \Sigma_g I_g L_g(m_i)$$

$$= \Sigma_g I_g \frac{\Sigma_i V_{gi}^2/m_i}{V_{gmin}}$$

$$= \Sigma_i \frac{1}{m_i} \Sigma_g \frac{I_g V_{gi}^2}{V_{gmin}} = \Sigma_i \frac{Z_i^2}{m_i}, \quad (6.3)$$

where $Z_i^2 = \Sigma_g I_g V_{gi}^2/V_{gmin}$. 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

$$\text{optimal } m_i^{**} = \frac{Z_i}{\sqrt{c_i}} \cdot \frac{C_f}{\Sigma Z_i\sqrt{c_i}} \quad \text{and} \quad (6.4)$$

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

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 $U_{gi} = V_{gi}\sqrt{c_i}/\Sigma V_{gi}\sqrt{c_i}$. 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

$$1 + L(m_i^{**}) = \sum_i \frac{Z_i^2}{m_i^{**}} = (\Sigma_i Z_i\sqrt{c_i})^2/C_f$$

$$= \frac{1}{C_f}\left[\sum_i \sqrt{\left(\Sigma_g \frac{I_g V_{gi}^2 c_i}{V_{gmin}}\right)}\right]^2. \quad (6.6)$$

When we accept (from 5.2) $V_{gmin} = (\Sigma V_{gi}\sqrt{c_i})^2/C_f$, we obtain a simpler form,

because $V_{gi}^2 c_i / V_{gmin} = (V_{gi}\sqrt{c_i}/\Sigma V_{gi}\sqrt{c_i})^2$.
Thus the jointly determined minimal loss function becomes

$$1 + L(m_i^{**}) = \{\sum_i \sqrt{[\sum_g I_g (V_{gi}\sqrt{c_i}/\sum_i V_{gi}\sqrt{c_i})^2]}\}^2.$$

$$= \{\sum_i \sqrt{(\sum_g I_g U_{gi}^2)}\} \qquad (6.7)$$

The minimal and optimal values may be unobtainable, due chiefly to the constraints $m_i^* \leq M_i$ (Section 5). In that case the above loss function overestimates the losses incurred over obtainable values of $V_{gmin}$. 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

$$\text{optimal } (m_i^{**}) = \frac{\sqrt{[\sum_g I_g U_{gi}^2]}}{\sum_i [\sum_g I_g U_{gi}^2]} \cdot \frac{C_f}{c_i}. \qquad (6.8)$$

These can be seen applied in 7F to the important and frequent conflict between allocations for weighted totals and comparisons of domains. Two examples are shown in Table 7.2. Note in the last column of 7.2B how encouragingly insensitive are the values of (6.8) 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, Ch.9). Another version (Yates, 1960, Cochran, 1963) uses $\sum_g I_g' \sum_{g_i} V_{gi}^2/m_i$, 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 $I_g' = I_g/V_{gmin}$ 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

$$F(m_i) = \sum_g I_{g_i} \sum V_{gi}^2/m_i V_{gmin} + \lambda \Sigma c_i m_i. \qquad (6.9)$$

With Lagrange Multipliers we also investigated two other loss functions: the product, $\Pi(1 + L_g)$, and the sum of the relative precisions, $[\Sigma(1 + L_g)^{-1}]^{-1}$. 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 $V + V_o$ for fixed $C_f$. In situations where $V_o$ is considerable, the actual loss should be modified to $L' = L/(1 + V_o/V_{min})$, 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, 196?, and Huddleston, 1970, were not the first]; but I do not find this approach useful.

## 7. $Var(\bar{y}) = \Sigma V_i^2/m_i + V_o$ IN SEVERAL APPLICATIONS

A) Stratified Sampling: $V_i = W_i S_i$

$$\Sigma\frac{(W_i S_i)^2}{m_i} - \Sigma\frac{W_i^2 S_i^2}{M_i} \qquad \text{opt } m_i^* = \frac{W_i S_i}{\sqrt{c_i}}$$

B) Multistage Random Selection of Equal Clusters:

2 stages $\dfrac{S_a^2 - S_b^2/B}{a} + \dfrac{S_b^2}{ab} - \dfrac{S_a^2}{A}$

$$\text{opt } b^* = \sqrt{(\frac{c_a}{c}\frac{S_b^2}{S_a^2 - S_b^2/B})}$$

$m_1 = a$, $m_2 = ab$, $m_3 = abc$

3 stages $\dfrac{S_a^2 - S_b^2/B}{a} + \dfrac{S_b^2 - S_c^2}{ab} + \dfrac{S_c^2}{abc} - \dfrac{S_a^2}{A}$

C) Two-Phase Sampling: Cost $= \Sigma c_i m_i + c_L n_L + C_o$

for Stratification:

$$\Sigma\frac{(W_i S_i)^2}{m_i} + \frac{\Sigma W_i(\bar{Y}_i - \bar{Y})^2}{n_L} - \frac{\Sigma(W_i S_i)^2}{M_i} + \frac{\Sigma W_h(\bar{Y}_i - \bar{Y})^2}{\Sigma M_i}$$

$= (\Sigma W_i S_i)^2/\Sigma m_i + \Sigma W_i(\bar{Y}_i - \bar{Y})^2/n_L$ when $m_i \propto W_i S_i$

for Regression: $\dfrac{S^2(1 - R^2)}{\Sigma m_i} + \dfrac{R^2 S^2}{n_L}$

D) Subsampling (1-P)m/k of Nonresponses:

Cost $= (c_o/P + c_p)Pm + c_q(1-P)m/k$

$$\frac{P^2 S_p^2}{Pm} + \frac{(1 - P)^2 S_q^2}{(1 - P)m/k}$$

$$\text{opt } k^* = \frac{S_p}{S_q}\left[\frac{c_q}{c_o/P + c_p}\right]^{1/2}$$

E) Weights in Estimation: $\Sigma W_i = \Sigma W_i^* = 1$

$V^2 = \Sigma W_i \sigma_i^2 \qquad \text{opt } W_i^* \propto 1/\sigma_i^2 \qquad V_{min} = 1/\Sigma 1/\sigma_i^2$

$1 + L = V^2/V_{min}^2 = \Sigma W_i^2/W_i^*$

$$L = \Sigma W_i^* (\frac{W_i}{W_i^*} - 1)^2$$

### 7F. Allocation Conflict Between Totals and Independent Domains

Serious conflict often exists between reducing the variance for the combined mean $\Sigma W_i \bar{y}_i$, and equal precision desired for the means $\bar{y}_i$ of H independent domains that differ greatly in relative sizes $W_i (\Sigma W_i = 1)$. 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 $V_c = \Sigma W_i^2 S_i^2 / m_i$ is minimal when the optimal $m^*_{ci} \propto W_i S_i / \sqrt{c_i}$. 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 <u>average domain variance</u>

$V_d = (\Sigma S_i^2 / m_i)/H^2$ for the variance of $\Sigma \bar{y}_i / H$. 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, their absence for the domain means. Thus the loss function (2.3) for the combined mean, due to allocation $m_i \propto S_i / \sqrt{c_i}$, has the departures $k_{ci} = m^*_i / m_i = W_i$, and the weights $U_{ci} \propto W_i S_i \sqrt{c_i}$. The loss function for the average domain means, due to allocations $m_i \propto W_i S_i / \sqrt{c_i}$, 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 $m^*_i \leq M_i$ in all domains. We shall also neglect effects of the constants $V_o$ and $C_o$ on the loss functions.

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

In Table 7.1 for loss functions $(1 + L_c)$ the minimal value 1 appears with $m_i \propto W_i$ for $\Sigma W_i \bar{y}_i$, and with $m_i \propto 1/H$ for $\Sigma \bar{y}_i / H$. 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, $H^2 \mathrm{Var}(W_i)$.

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

$$\begin{aligned} 1 + L_j(m_i) &= I_c m \Sigma W_i^2 / m_i + I_d m H^{-2} \Sigma 1/m_i \\ &= m\Sigma \left[ I_c W_i^2 + I_d H^{-2} \right] / m_i \\ &= mH^{-2} \Sigma_i [I_c D_i^2 + I_d] / m_i \\ &= mN^{-2} \Sigma_i [I_c N_i^2 + I_d \bar{N}^2] / m_i \\ &= m\Sigma_i t_i^2 / m_i . \end{aligned} \quad (7.23)$$

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.

We find the joint optimal allocations $m^{**}_i = mt_i / \Sigma t_i$ where the

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

The $m^{**}_i$ may be found simply with (5.3) but also as an illustration of (6.8).

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

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

When we use the multipurpose optimal $m^{**}_i \propto t_i$ we can determine the loss functions (1 + L) incurred for the variances of $\Sigma W_i \bar{y}_i$ and $\Sigma \bar{y}_h / H$; we use (6.1) or (6.2) with

$$k_{ci} \propto V_{ci} / V'_i \propto W_i / t_i$$

and $\quad k_{di} \propto V_{di} / V'_i \propto 1/Ht_i \quad$ respectively.

These (1 + L) are shown on the bottom row of Table 7.1. 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 7.2. 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.

Conflict of Combined Mean ($\Sigma W_i \bar{y}_i$) and Average Domain Mean ($\Sigma \bar{y}_i / H$)

($S_i^2$ and $c_i$ are assumed constant and omitted.)

Table 7.1 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 | | |
|---|---|---|---|
| | $\Sigma W_i \bar{y}_i$ | $\Sigma \bar{y}_i / H$ | $I_c \Sigma W_i \bar{y}_i + I_d \Sigma \bar{y}_i / H$ |
| $(1 + L) = mV^2$ | $m \Sigma W_i^2 / m_i$ | $m H^{-2} \Sigma 1 / m_i$ | $m \Sigma t_i^2 / m_i$ |
| Allocation of $m_i$ | | | |
| $mW_i$ | 1 | $H^{-2} \Sigma 1 / W_i$ | $I_c + I_d H^{-2} \Sigma 1 / W_i$ |
| $m/H$ | $H \Sigma W_i^2 = 1 + C_w^2$ | 1 | $I_c H \Sigma W_i^2 + I_d H^{-1}$ |
| $mt_i / \Sigma t_i$ | $(\Sigma W_i^2 / t_i)(\Sigma t_i)$ | $H^{-2}(\Sigma 1/t_i)(\Sigma t_i)$ | $(\Sigma t_i)^2$ |

Table 7.2 Loss Functions (1 + L) for Two Populations

| Allocations $m_i$ | (A) (1 + L) for $W_1/W_2 = 4$: $\Sigma W_i \bar{y}_i$ | $\Sigma \bar{y}_i / 2$ | Joint | (B) (1 + L) for 133 countries: 0.2 to 100 mm: $\Sigma W_i \bar{y}_i$ | $\Sigma \bar{y}_i / 133$ | Joint with weights 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 | |
| $\propto \sqrt{(0.5W_i^2 + H^{-2})}$ | | | | 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 ($W_1 = 0.8$ and $W_2 = 0.2$); note that the allocation $m_i^* \propto \sqrt{W_i}$ 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.