Design Optimization With Multiple Objectives

James R. Chromy, Research Triangle Institute
Box 12194, Research Triangle Park, NC 27709

A general problem in sampling theory and survey design is how to best allocate a sample in order to minimize cost subject to a maximum variance restriction.

Neyman (1934) considered sample allocation applied to stratified sampling. In this case, the variance function for the estimated mean is easily partitioned into a sum of population-based constants divided by stratum sample sizes plus a term that does not depend on the stratum sample sizes. Neyman's model can be put in the form of a single variance function.

$$V(Y)_{st} = \sum_h w^2(h)\, S^2(h)/n(h) - \sum_h w^2(h)\, S^2(h)/N(h) = \sum_h w^2(h) S^2(h)/n(h) + V_o$$

The variance inequality can be put in the form $\sum_h V(h)/n(h) \leq V^*$ where V^* = required variance - V_o. The sample size subject to the variance constraint is minimized when the stratum sample sizes are proportional to the square root of the complete variance components, V(h). Neyman applied a calculus approach to obtain the optimum solution.

Stuart (1954) utilized the Cauchy-Schwarz inequality to obtain a simple approach to sample allocation with a variable cost function. In the case of a single variance constraint, either the variance or the cost may be fixed. The form of the solution for the sample sizes is the same in either case; only the constant of proportionality changes as a function of the particular constraint imposed.

Optimum allocation problems arise in sample designs other than stratified sampling. Unique solutions are available in the textbooks and literature for multistage sampling, cluster sampling, multi-phase sampling, and combinations thereof. Kish (1974) in a paper on optimal and proximal multipurpose allocation notes that proper isolation of population parameters and sample size parameters allows variance functions for most sample allocation problems to be expressed in a common form. Moreover, variances of common nonlinear estimates can generally be put in a corresponding form using appropriate linear approximation methods.

In practice, very few surveys are ever conducted with the sole objectives of minimizing the variance of a single estimate. Multiple estimates are most commonly needed for different domains or for different measures on the same sampling units.

Sample allocation based on a single variance constraint model usually involves a process of prioritizing different estimates and selecting one on which to base the design. A more appealing strategy is to consider several different estimates simultaneously or to consider classifying the different estimates according to their variance properties and selecting the typical variance model from each class. One then proceeds to find a minimum cost design which meets the variance requirements for all of the selected estimates. Before describing the mechanics of this approach, I wish to discuss an alternate approach which starts with the concept of a fixed cost.

The fixed cost approach is advocated by Kish (1974). Let me quote his view before I state mine: "Furthermore, I consider fixing C_f (total cost bound) more practical than trying to fix values for a set of Vg (variance bounds) and then to minimize C_f. This problem seems to have been solved with convex programming on several separate occasions, but I do not find this approach useful."

I agree with Dr. Kish that there is little purpose in trying to understand a $2 social problem with a $10 survey; you can multiply these dollar figures by a million to reflect the kind of survey research often conducted at the national level. However, as a matter of principle, I consider it unnecessary and wasteful to design and conduct a larger and more costly survey than the one that obtained the needed precision. Critical resources saved by controlling the size of the survey can be devoted to increasing the quality of the data or to enhancing the utilization of the data through more extensive analysis. The cost limitations approach to survey optimization is perhaps well-founded in practical experience. It may be observed that clients rarely have the resources to fund the ideal survey. Blind adherence to a cost limitations strategy could however also lead to the execution of a cost-feasible survey that lacked the capability of satisfying any of the research objectives. Since multiple variance constraints often lead to crossed purposes in sample allocation, a so-called "compromise allocation" between two opposing objectives can produce a design which meets neither objective. A client would be better counseled to select one objective and do it well or to seek the additional resources required to meet both objectives before starting the survey. Waters and Chester (1987) show graphically for a two-variable problem that the solution that minimizes cost and satisfies multiple constraints may not appear to be an obvious compromise.

Cost limitations can be introduced into the design optimization process based on variance constraints by deciding which variance constraints can be relaxed while still preserving the major objectives of the study. To the extent that the variance models are good approximations to reality, this method of negotiating the total cost of a survey ensures that the analyses that can be supported by the information content of the resulting survey data base are consistent with pre-survey expectations.

Having noted my disagreement with Dr. Kish's preference for a cost constraint approach, I

wish to acknowledge that the form of the solution for Dr. Kish's proximal multipurpose allocation is identical to the form of the solution obtained by starting with variance constraints. Indeed, Dr. Kish's result triggered my exploration of this approach after becoming familiar with Kuhn-Tucker theory. Both solutions substitute a weighted sum of population variance components for a single population variance component in the single constraint solution. The values of the weighting function are motivated differently under the two approaches.

As Kish also pointed out, the solution to the multiple variance constraint problem is not new. Exact solutions utilizing convex programming are discussed by Huddleston, Claypool, and Hocking (1970). Earlier graphic methods for small problems were developed by Dalenius (1957). In spite of the availability of exact methods, many approximate methods suitable for hand calculation are discussed in Cochran (1977). I am aware that the U.S. Department of Agriculture routinely uses the procedures of Huddleston et al in designing many of their surveys. More recently, James Bethel (1985) reported on an updated algorithm used at USDA. Statisticians at RTI have used the algorithm discussed in this paper for several survey designs including a sample allocation for a medical provider record check survey (Folsom, Chromy, and Williams, 1979), a sample design for a youth attitude tracking study (Mason and Sweetland, 1983), and an evaluation of alternate designs for future medical expenditure surveys (Cox and Folsom, 1984).

The algorithm I will be discussing is based on Kuhn-Tucker theory (Kuhn and Tucker, 1951; Simmons, 1975). This theory is used by Hughes and Rao (1979) for optimum allocation with some selected types of multiple constraints, mostly for cases involving several linear constraints and a single variance constraint. Kuhn-Tucker theory is also applied by Thompson (1962) to variance component estimation. Both of these applications involve a quick search of corner points under a set of linear constraints. Bethel's (1985) algorithm also applies Kuhn-Tucker theory.

The single variance constraint variance and cost models can be expressed as follows:

$$\mathrm{Var}(\hat{\mu}) = V + V_o = \sum_{h=1}^{H} V(h)/x(h) + V_0$$

where

V_0 = portion of variance that is not a function of sample size (may be positive or negative);

V(h) = a non-negative function of population variance and weighting factors associated with the hth term of the variance; and

x(h) = a sample size or a product of sample sizes (in a multistage design) associated with the hth term of the variance expression.

Since $\mathrm{Var}(\hat{\mu}) - V_0 = V$, it is possible to consider constraining $\mathrm{Var}(\hat{\mu})$ by considering only

$$V = \sum_{h=1}^{H} V(h)/x(h).$$

The most commonly used cost model can be put in the form:

$$\mathrm{Cost} = C + C_0 = \sum_{h=1}^{H} C(h)x(h) + C_0$$

where

C_0 = fixed cost component (not a function of sample size);

C(h) = a positive cost component associated with one unit of x(h); and

x(h) = a function of sample sizes as defined above.

Since Cost - C_0 = C, cost may be minimized in terms of the x(h) by considering only

$$C = \sum_{h=1}^{H} C(h)x(h).$$

The steps for solving the single variance constraint problem may be listed as follows:

1. Require that $V \leq V^*$ where V^* is specified by external requirements (e.g., based on a planned statistical hypothesis test or on requirements related to the length of confidence intervals).

2. Given V^*, choose x(h), h = 1,2,..., H, to minimize C, i.e.,

   $$\text{Minimize } C = \sum_{h=1}^{H} C(h)x(h)$$

   subject to

   (1) $\sum_{h=1}^{H} V(h)/x(h) \leq V^*$, and

   (2) $x(h) \geq 0$ for h = 1,2,..., H.

3. Solve by treating the variance constraint as an equality constraint and applying Lagrange multipliers approach or Cauchy-Schwarz solutions. This yields

$$x^*(h) = [\lambda\, V(h)/C(h)]^{1/2}$$

$$\text{with } \lambda = \left[\sum_{h=1}^{H} \sqrt{V(h)C(h)}\,/\,V^*\right]^2$$

The multiple constraint problem is stated by adding an index, k, to the variance function.

$$V(k) = \sum_{h=1}^{H} V(k,h)/x(h)$$

The constraints take the form

$$V(k) \leq V^*(k) \text{ for } k = 1,2,\ldots,K.$$

The cost model remains unchanged. The problem is stated as

$$\text{Minimize } C = \sum_{h=1}^{H} C(h)x(h)$$

subject to

(1) $\sum_{h=1}^{H} V(k,h)/x(h) \leq V^*(k)$ for $k=1,2,\ldots,K$,

and

(2) $x(h) \geq 0$ for $h=1,2,\ldots, H$.

The steps for solving this form of the multiple constraints problem are as follows:

1. Specify constants.

   a. Cost components $C(h)$, $h=1,2,\ldots, H$.

   b. Variance components $V(k,h)$ $h=1,2,\ldots, H$ and $k=1,2,\ldots, K$.

   c. Variance bonds $V^*(k)$ $k=1,2,\ldots, K$.

   d. Check constraints on constants

      (1) $C(h) > 0$ for $h=1,2,\ldots, H$.

      (2) $V(k,h) \geq 0$ for $h=1,2,\ldots, H$ and $k=1,2,\ldots, K$ and

      (3) $\sum_{h=1}^{H} V(k,h) > 0$ for $k=1,2,\ldots, K$.

      (4) $V^*(k) > 0$.

2. Set $\lambda_i(k) = 1$ for $k=1,2,\ldots, K$ for the first iteration ($i=1$).

3. $$\text{Minimize } F(\underset{\sim}{x}) = \sum_{h=1}^{H} C(h)x(h) + \sum_{k=1}^{K} \lambda_i(k) \sum_{h=1}^{H} V(k,h)/x(h)$$

   Setting $\partial F(x)/\partial x(h) = 0$ for each $x(h)$ yields

   $$x_i(h) = \left[\sum_{k=1}^{K} \lambda_i(k)\, V(k,h)/C(h)\right]^{1/2}$$

4. Compute resulting variances and adjust $\lambda_i(k)$ values

   $$V_i(k) = \sum_{h=1}^{H} [V(k,h)/x_i(h)]$$

   $$\lambda_{i+1}(k) = \lambda_i(k)[V_i(k)/V^*(k)]^2$$

5. Increment i and repeat steps 3 and 4 until the Kuhn-Tucker conditions are satisfied. Note that three sets of conditions are met at each step by the form of the solution method.

   (1) $\partial F(x)/\partial x(h) = 0$ for $h=1,2,\ldots, H$.

   (2) $\lambda_i(k) \geq 0$ for $k=1,2,\ldots, K$.

   (3) $x_i(h) \geq 0$ for $h=1,2,\ldots, H$.

Kuhn-Tucker theory provides that if in addition to the above,

(1) $V_i(k) \leq V^*$ for $k=1,2,\ldots, K$, and

(2) $\lambda_i(k)\,[V^*(k) - V_i(k)] = 0$ for $k=1,2,\ldots, K$,

then the solution obtained is one that minimizes cost subject to the constraints.

Note in step 3 that the form of the solution for $x(h)$ is identical to the form of the form of the solution in Kish's proximal solution method. In this case the weights, $\lambda_i(k)$, derive from the iterative adjustment method which leads to the Kuhn-Tucker conditions for an optimum. In practice, many of the weights will be zero, identifying the slack constraints. One or more nonzero weights will identify the constraints that determine the solution.

If $K=1$ (the single constraint problem) or if $V(k,h) > 0$ for only one value k for all $h=1,2,\ldots, H$, then an exact solution is reached after 2 iterations. Convergence can be obtained more quickly by zeroing out those λ that appear to be converging toward 0. This can be done by operator intervention on an interactive computing system or within the program itself.

It is sometimes necessary to constrain certain sample sizes. Many of these constraints can be stated in terms of ratios. For example in double sampling, the second phase sample is a subsample of the first phase sample. If we designate them by $x(2)$ and $x(1)$, respectively, the appropriate ratio constraint is to require that the ratio of $x(2)$ to $x(1)$ be less than unity.

In a three-stage sample, $x(3)$ could represent students sampled and $x(2)$ could represent schools sampled. From practical considerations, we may wish to limit the number of students

sampled per school to be 30 or less, based on the size of classrooms usually available for a special testing program. This constraint could be stated as $x(3)/x(2) \leq 30$. Similarly if $x(1)$ represented the number of districts sampled, we might want to sample at least 2 schools per district. To put this in the proper form, we would have to require that the number of districts per school be less than 1/2, i.e., $x(1)/x(2) \leq .5$.

The general form of the problem with sample size ratio constraints involves adding an index j to define J constraints of the form $R(j) \leq R^*(j)$ where

$$R(j) = x[h(j,1)]/x[h(j,2)]$$

The problem statement is then

$$\text{Minimize } C = \sum_{h=1}^{H} C(h)x(h)$$

subject to

(1) $\sum_{h=1}^{H} V(k,h)/x(h) \leq V^*(k)$ for $k = 1,2,\ldots,K$,

(2) $x(h) \geq 0$ for $h = 1,2,\ldots, H$, and

(3) $R(j) \leq R^*(j)$ for $j = 1,2,\ldots J$.

The steps to obtaining a solution for this expanded problem are:

1. Specify constants.

   a. Cost components $C(h)$, $h=1,2,\ldots, H$.

   b. Variance components $V(k,h)$ $h=1,2,\ldots, H$ and $k=1,2,\ldots, K$.

   c. Variance bonds $V^*(k)$ $k = 1,2,\ldots, K$.

   d. Ratio constraints require specifications of a numerator index, $h(j,1)$; a denominator index, $h(j,2)$; and the bound $R^*(j)$.

   e. Check constraints on constants (first 4 steps as in the original problem)

      (5) $1 \leq h(j,1) \leq H$, $1 \leq h(j,2) \leq H$, and both are integers.

      (6) $R^*(j) > 0$.

2. Set $\lambda_i(k)=1$ for $k=1,2,\ldots, K$ for the first first iteration ($i=1$).

3. Set $\gamma_i(j) = 0$ for $j=1,2,\ldots, J$ for the first iteration ($i=1$).

4. Minimize $F(\underset{\sim}{x}) = \sum_{h=1}^{H} C(h)$

$$+ \sum_{k=1}^{K} \lambda i(k) \sum_{h=1}^{H} V(k,h)/x(h)$$

$$+ \sum_{j=1}^{J} \gamma_i(j)\, x[h(j,1)]/x[h(j,2)].$$

Setting $\partial F(x)/\partial x(h) = 0$ for each $x(h)$ yields

$$x_i^2(h) = \left\{ \sum_{k=1}^{K} \lambda_i(k)\, V(k,h) + \sum_{j \ni h(j,2)=h} \gamma_i(j)\, x[h(j,1)] \right\} \Big/ \left\{ C(H) + \sum_{j \ni h(j,1)=h} \gamma_i(j)/x[h(j,2)] \right\}$$

5. Compute resulting variances and adjust $\lambda_i(k)$ values

$$V_i(k) = \sum_{h=1}^{H} [V(k,h)/x_i(h)]$$

$$\lambda_{i+1}(k) = \lambda_i(k)[V_i(k)/V^*(k)]^2.$$

6. Compute $R_i(j) = x_i[h(j,1)/x_i[h(j,2)]$

   (a) If $\gamma_i(j) = 0$ and $R(j) \leq R^*(j)$, no action is required.

   (b) If $\gamma_i(j) = 0$, and $R(j) > R^*(j)$, initialize $\gamma_i(j) = 1$.

   c. If $\gamma_i(j) > 0$, adjust $\gamma_i(j)$ as

$$\gamma_{i+1}(j) = \gamma_i(j)\, [R_i(j)/R^*(j)]^2.$$

7. Increment i and repeat steps 3 through 5 until the Kuhn-Tucker conditions are satisfied.

A simple example in three-stage sampling illustrates the method and the use of notation. Suppose n(1), n(2), and n(3) represent the number of first-stage units, second-stage units per first-stage unit, and third-stage units per second-stage unit, respectively. Then

$x(1) = n(1)$
$x(2) = n(1)\, n(2)$
$x(3) = n(1)\, n(2)\, n(3)$

Example 1:

Constants are specified to define the problem as follows:

a. Cost components: $C(h) = 1$, for $h = 1,2,3$.

b. Variance components: $V(k,h)$

	h=1	h=2	h=3
k=1	.01	.14	.85
k=2	.10	.10	.80
k=3	.05	.05	.90

c. Variance bounds

$V^*(1) = .05$, $V^*(2) = .075$, $V^*(3) = .05$.

Partial solutions based on applying each constraint, k, individually are:

k	x(1)	x(2)	x(3)	λ	C
1	2.79	10.45	25.74	779.65	38.98
2	6.44	6.44	18.21	414.47	31.08
3	6.24	6.24	26.48	779.41	38.98

A combined solution which satisfies all the constraints is obtained after 31 iterations.

x(1)	x(2)	x(3)	C
4.83	8.83	26.49	40.15

Values of $\lambda_i(k)$ and $V_i(k)$ for i=31 are shown below. The variance bounds $V^*(k)$ are shown for comparison with computed variances.

k	$\lambda_{31}(k)$	$V_{31}(k)$	$V^*(k)$
1	420.46	.05000	.05
2	0.00	.06222	.075
3	382.45	.04998	.05

Note that constraint 2 (k=2) is a slack constraint; the multiplier $\lambda_{31}(2)$ is zero and the computed variance, $V_{31}(2)$, is strictly less than the specified constraint value, $V^*(2)$.

Example 2:

A second example adds ratio-type constraints. Suppose we require that $n(3) \geq 2$ and $n(2) \geq 2$. This translates to $x(2)/x(3) \leq .5$ and $x(1)/x(2) \leq .5$. The same constants are used to specify the problem as in example 1 except that additional ratio constraints, $R^*(j)$, are added

	h(j,1)	h(j,2)	$R^*(j)$
j=1	2	3	.5
j=2	1	2	.5

A combined solution is obtained after 228 iterations

x(1)	x(2)	x(3)	C
4.50	9.00	27.00	40.50

As expected, adding constraints increases costs. Values of $\lambda_i(k)$, $V_i(k)$ compared to $V^*(k)$, $\gamma_i(j)$, and $R_i(j)$ compared to $R^*(j)$ are shown below:

k	$\lambda_{228}(k)$	$V_{228k}(k)$	$V^*(k)$
1	0.46	.04926	.05
2	0.00	.06296	.075
3	809.55	.05000	.05

j	$\gamma_{228}(j)$	$R_{228}(j)$	$R^*(j)$
1	0.46	.50000	.50
2	0.00	.33333	.50

In this case, the design is determined by the variance constraints with k = 1 and k = 3 and by the ratio constraint with j = 1.

Acknowledgement

The author wishes to acknowledge the assistance of Dr. Robert E. Mason, Research Triangle Institute, for testing the algorithm on applied problems and for reviewing drafts of this manuscript.

References

Bethel, James (1985), "An Optimum Allocation Algorithm for Multivariate Surveys," presented at Annual Meeting of the American Statistical Association.

Chaterjee, S. (1966), "A Programming Algorithm and its Statistical Applications," O.N.R. Technical report 1, Department of Statistics, Harvard University, Cambridge.

Cochran, William G. (1977). Sampling Techniques, Third Edition, New York: John Wiley & Sons.

Cox, Brenda G., and Folsom, Ralph E. (1984), "Evaluation of Alternate Designs for a Future NMCUES," Paper presented at he Joint Statistical Meetings of the American Statistical Association.

Dalenius, T. (1957), Sampling, in Sweden, Contributions to the Methods and Theories of Sample Survey Practice, Almquist and Wicksell, Stockholm.

Folsom, Ralph E. Jr., Chromy, James R., and Williams, Rick L. (1979). "Optimum Allocation of a Medical Care Provider Record Check Survey: An Application of Survey Cost Minimization Subject to Multiple Variance Constraints," Paper presented at the Joint National Meetings of the Institute of Management Science and the Operations Research Society.

Hartley, H. O., and Hocking, R. (1963), "Convex Programming by Tangential Approximation," Management Science,9, 600-612.

Huddleston, H. F., Claypool, P. L., and Hocking, R. R., (1970), "Optimum Sample Allocation to Strata Using Convex Programming," Applied Statistics, 19, 273-278.

Hughes, Edward, and Rao, J. N. K., (1979), "Some Problems of Optimal Allocation in Sample Surveys Involving Inequality Constraints," Communications in Statistics - Theory and Methods, A8 (15), 1551-1574.

Kish, Leslie (1974), "Optimal and Proximal Multipurpose Allocation," Proceedings of the Social Statistics Section, American Statistical Association, 111-118.

Kokan, A. R. (1963), "Optimum Allocation in Multivariance Surveys," Journal of the Royal Statistical Society, A126, 557-565.

Kuhn, H., and Tucker, A. W. (1951), "Nonlinear Programming," in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, California, 481-492.

Mason, R. E., and Sweetland, S. S., (1983), "Sampling Design, Youth Attitude Tracking Study II," Research Triangle Institute, Report No. RTI/2622/02-01W.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," Journal of the Royal Statistical Society, 97, 558-606.

Simmons, Donald M. (1975), Nonlinear Programming for Operations Research. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Stuart, A. (1954), "A Simple Presentation of Optimum Sampling Results," Journal of the Royal Statistical Society, B16-, 239-289.

Waters, James R., and Chester Alexander J. (1987), "Optimum Allocation in Multivariate, Two-Stage Sampling Designs," The American Statistician, Vol 41, Number 1, pp. 4650.

Yates, F. (1960), Sampling Methods for Census and Surveys, Third Edition, Charles Griffin and Co., London.

Zukhovitsky, S. I., and Avdeyeva, L. I. (1966). Linear and Convex Programming, W. B. Sanders, Philadelphia.