MATHEMATICAL PROGRAMMING FOR SAMPLE DESIGN AND ALLOCATION PROBLEMS

James L. Green
Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Mathematical programming; Linear programming; Design effects; Two-wave sampling; Three-dimensional stratification

1. Introduction

Survey sample design in general and sample allocation problems in particular can benefit from a mathematical programming formulation, especially when operational or sample size constraints lead away from straightforward or closed-form solutions. This paper presents an introduction to this approach and two specific applications. The paper concludes with some discussion of more complicated problems and suggestions for future applications and research.

2. Sample Design

Sampling statisticians are principally concerned with the design of efficient samples. Sample efficiency is measured either by cost or precision, given the other, and the final sample design often reflects an attempt to strike some compromise between the two. An explicit solution is available for practically every sample design, at least in the context of a single variable of interest (Cochran, Hansen et. al., Kish, 1965). These solutions either solve for the sample size required for specified precision or indicate the precision offered by a specific sample size. These solutions can become more complicated with additional constraints (e.g., on stratum specific sample sizes) or variables of interest.

The precision of a sample design is usually measured by the sampling variance of the statistic of interest. Variances are typically controlled by stratification and clustering and the effect of these techniques are measurable or estimable. The variance implications of other design features, such as a deviation from proportional allocation or the introduction of differential weighting effects are also measurable (Kish, 1992).

The cost of a proposed sample design can be measured naively by sample size or through a more complex cost function if the appropriate data is available. When such data is available at the stratum level along with variances, optimal allocation formulas may be used (Cochran.) These formulae essentially assume a single variable of interest.

Given the truly multi-variate nature of most sample surveys, other techniques are required for more complicated and efficient design. Mathematical programming has been used to solve the multi-variate sample design optimization problem (Bethel, Leaver et. al., Valliant and Gentle) and is also appropriate for a wide range of other problems encountered in sample design (see section 5 for an overview.)

3. Mathematical Programming

Mathematical programming problems have the following three characteristics:

1. Decision variables.
2. An objective function.
3. Constraints on the decision variables.

In general, the decision variables are allowed to take on any values as needed, subject to certain constraints, in order to solve all other aspects of the mathematical programming problem. The objective function is a function of the decision variables and can be either linear on nonlinear. The nature of the problem often requires that the objective function be maximized, minimized or set to a particular value. The constraints are often explicit constraints on the decision variables themselves or implicit constraints on other linear or non-linear functions of the decision variables. Sometimes the decision variables are subject to certain reality or practicality constraints. For example, sample sizes and sampling rates must be > 0. A general linear programming problem might be expressed as follows:

Decision variables:	x_i
Objective function:	$MIN(\sum K_i x_i)$
Constraints:	$L_i \leq x_i \leq H_i$
	$\sum C_i x_i \leq Y$

4. Two Specific Applications

4.1 Two-Wave Sample Allocation Problem

The Substance Abuse and Mental Health Services Administration's (SAMHSA) Alcohol and Drug Services Survey (ADSS) used a national probability sample of drug treatment facilities. The sampling frame for this survey was a list of drug treatment facilities maintained by SAMHSA. Stratification information was available on the frame, however the information was known to be out-of-date due to considerable changes in the treatment facility population and included missing values that made it impossible to assign some facilities to a specific

sampling stratum. Seven sampling strata were used, with six strata being of analytic interest and the seventh sampling stratum containing the facilities with missing stratification information.

Facilities were sampled in two waves. The first wave was released in order to obtain sample-based estimates of the sampling stratum – actual stratum transition matrix. The second wave sample allocation was based on the transition matrix and the actual stratum target sample sizes.

This was handled as a linear programming problem (specifically, a transportation problem.) The problem was expressed in the following linear programming notation:

Minimize: $\sum_{i=1}^{I} n_i''$

Subject to: $0 \le n_i'' \le N_i''$;

$n_j'' \ge T_j''$

Where: $n_j'' = \sum_{i=1}^{I} n_i'' p_{ij}'$

n_i'' is the second wave sample size in sampling stratum i

N_i'' is the population size in sampling stratum i after wave 1

n_j'' is the expected second wave sample size in actual stratum j

T_j'' is the required sample size for wave 2 in actual stratum j

p_{ij}' is the wave 1 percent of sampled units in sampling stratum i that belong in actual stratum j.

Table 1 below gives the solutions we obtained for this problem. Note that we could not reach the target sample size for actual stratum 1 due to the maximum workload limitation (a value less than the population size) and the poor relationship between this actual stratum and any of the sampling strata. Since we could not obtain a feasible solution, we modified the results in Table 1 to reach results that were more or less acceptable.

4.2 Three-Dimensional Stratification Problem

The National Center for Educational Statistics' (NCES) Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) will use a national probability sample of children born in the year 2001. The sampling frame for this survey will be birth registrations available through the National Center for Health Statistics (NCHS.) Children will be sampled throughout the year 2001 and on a flow basis as NCHS receives the registered births. Detailed data is available on the birth certificate, including mother's and father's race/ethnicity, the child's birth weight and plurality (single birth, twins, triplets etc.) all of which represent specific analytic domains with varying numbers of levels and specific precision requirements. Similar detailed data was available for previous years and was used to estimate the 2001 population. Using each domain as an independent stratification variable, the sample allocation problem was treated as a multi-dimensional stratification problem. Figure 1 provides an illustration of our three dimensional problem. A birth in the year 2001 could fall into any one of the thirty cells that make up this cube, yet contribute to one level within each of three domains. Overall sampling rates were required for each of the 30 cells and were calculated given target actual sample sizes and expected population counts.

Table 1. Two-wave sample allocation – solution

Sampling stratum (i)	Maximum workload (N_i'')	Sample size (n_i'')	Actual stratum (j)					
			1	2	3	4	5	6
1	603	603.0	0.58	0.04	0.01	0.01	0.07	0.30
2	600	600.0	0	0.82	0.01	0	0.01	0.15
3	463	463.0	0	0	0.92	0	0.07	0.01
4	598	482.9	0	0	0.01	0.53	0.43	0.03
5	1,025	819.8	0	0.01	0.03	0.09	0.84	0.03
6	595	595.0	0.01	0.03	0.01	0.02	0.27	0.66
7	807	807.0	0.01	0.13	0.03	0.17	0.57	0.08
	Actual (n_j'')	4370.7	363.8	647.1	497.6	484.8	1597.5	771.9
	TARGET (T_j'')	2534	374	374	374	374	664	374

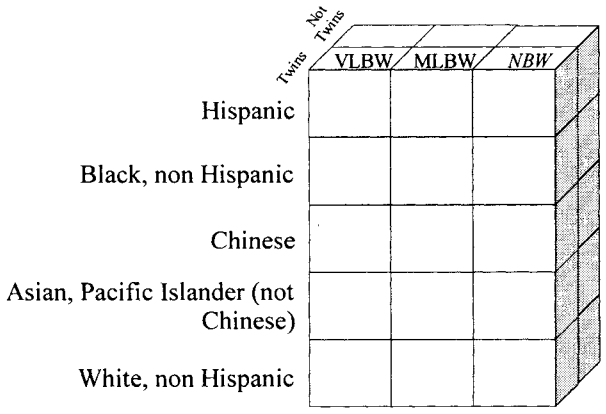

Figure 1: Three Dimensional Stratification

This was handled as a mathematical programming problem. The problem was expressed in the following mathematical programming notation:

Minimize: $\sum_{H}\sum_{I}\sum_{J} n_{hij}$

Subject to: $0 < n_{hij} \leq N_{hij}$;

$$\frac{\sum_{I}\sum_{J} n_{hij}}{d_h} \geq t_h \text{ etc.}$$

where

n_{hij} is the actual sample size in cell hij

N_{hij} is the population size in cell hij

t_h, t_i, t_j are the target effective sample sizes of levels h, i, j in domains H, I, J; and

$d_h = \frac{n_h}{N_h^2}\sum_{I}\sum_{J}\frac{N_{hij}^2}{n_{hij}}$ etc. are the design effects for levels h, i, j in domains H, I, J.

Table 2 gives the actual sample size solutions we obtained for the thirty distinct sampling strata that result from crossing all levels of all domains. Table 3 gives the actual and effective (i.e., adjusted for the differential weighting effects) sample sizes by level of domain. Note that this problem could be extended to practically any number of domains with any number of levels.

Table 2. Three dimensional stratification – solution by cell (n_{hij})

Race/ethnicity X birth weight	Twins	Not twins	Total
Hispanic, VLBW	36.6	182.6	219.3
Hispanic, MLBW	84.9	177.0	261.9
Hispanic, *NBW*	88.7	1,397.7	1,486.4
Black, VLBW	90.3	434.6	524.9
Black, MLBW	129.6	320.9	450.5
Black, *NBW*	91.8	1,221.2	1,313.0
Chinese, VLBW	1.4	5.4	6.8
Chinese, MLBW	5.7	19.8	25.6
Chinese, *NBW*	5.4	502.9	508.3
API (not Chinese), VLBW	7.1	40.7	47.8
API (not Chinese), MLBW	20.3	81.6	101.9
API (not Chinese), *NBW*	17.8	1,184.9	1,202.7
White, VLBW	180.2	615.2	795.4
White, MLBW	428.0	587.8	1,015.8
White, *NBW*	481.3	3,966.1	4,447.4
Total	1,669.3	10,738.4	12,407.7

Table 3. Three dimensional stratification – solution by level of domain

Analytic subgroup	Actual Wave 1 completes (n_{hij})	Weighting effect (d_h etc.)	Effective Wave 1 completes
Hispanic	1,968	1.24	1,590
Black, non-Hispanic	2,288	1.44	1,590
Chinese	541	1.00	539
Asian, Pacific Islander (not Chinese)	1,352	1.02	1,322
White, non-Hispanic	6,259	1.37	4,572
Very low birth weight (VLBW)	1,594	1.00	1,590
Moderately low birth weight (MLBW)	1,856	1.17	1,590
Normal birth weight (NBW)	8,958	1.25	7,156
Twins	1,669	1.05	1,590
Single births and other non-twins	10,738	1.38	7,810
Total	12,408	1.51	8,190

5. Other Applications

Mathematical programming solutions have been developed for a variety of sample design problems, including raking, controlled selection, multi-way stratification, overlap control and multivariate sample design. Table 4 below provides a cross-reference of selected problems and references.

Raking, or iterative proportional fitting, has been treated as a linear programming problem (Arthanari and Dodge; Causey, Cox and Ernst), specifically as a transportation problem. Controlled selection and multi-way stratification have been treated as linear programming problems (Causey, Cox and Ernst; Rao and Nigam; Sitter and Skinner), with this approach offering a nice and previously difficult to obtain solution to the controlled selection problem. Overlap control has been treated as a linear programming problem, with the size of the problem growing rapidly with changes in strata and large numbers of units (Arthanari and Dodge; Causey, Cox and Ernst). An alternative linear programming procedure has been proposed to help reduce the size of the overlap control problem and may be reasonably optimal in practice (Ernst and Ikeda.) Multivariate sample design and optimization has been treated as a mathematical programming problem (Arthanari and Dodge; Bethel; Leaver et. al; Valliant and Gentle), with the nonlinearity due to precision related terms either in the objective function or in the constraining equations.

6. Suggested Applications and Research

Mathematical programming can offer solutions to atypical designs, where standard sample design formulae are inappropriate. For example, study directors often prefer to pick a small number of field test PSUs purposively, covering a number of desirable properties with each selected PSU. This purposive selection essentially eliminates the first stage selection probability, which usually ensures desirable sample design properties. For example, under PPS sampling, the first stage measure of size and the resulting probabilities can be used to reflect a PSU's concentration of a rare subgroup as well as achieve (at least roughly) equal within PSU workloads. In the absence of the first stage selection probability, a mathematical programming formulation can be used to calculate within PSU selection probabilities (e.g., in a two-stage design) while controlling the variability in within PSU workloads and also the deviation from proportional allocation to subgroups.

Mathematical programming approaches may also be useful in developing automatic PSU formation software and automatic stratification software. Both of these problems could be conceived as binary programming problems, given an appropriate objective function (i.e., cost or variance) and a number of constraints, all functions of the binary decision variables. Binary programming problems are notorious, however, for their tendency to become extremely large very quickly and therefore difficult or impossible to solve with most modern computers. The PSU formation and stratification problems are no exception to this rule, however some of the mathematical programming concepts can still prove useful in developing algorithms that are relatively optimal. Our work to date at Westat on automatic PSU formation software has been very encouraging and may be a future presentation topic.

Table 4. Mathematical programming problems in sample design and associated references

Application	Reference(s)						
	Arthanari & Dodge	Bethel	Causey, Cox & Ernst	Leaver et. al.	Rao & Nigam	Sitter & Skinner	Valliant & Gentle
Raking	X		X				
Controlled Selection/Multi-way Stratification			X		X	X	
Overlap Control	X		X				
Multivariate Sample Optimization	X	X		X			X

7. References

Arthanari, T.S., and Dodge, Y. (1993). *Mathematical Programming in Statistics*. Wiley, New York.

Bethel, James. (1989). "Sample Allocation in Multivariate Surveys." Survey Methodology, 15-1, 47-57.

Causey, B.D., Cox, L. H., and Ernst, L.R. (1985). "Applications of Transportation Theory to Statistical Problems." Journal of the American Statistical Association, 80, 903-909.

Cochran, William G. (1977). *Sampling Techniques*. Wiley, New York.

Ernst, L.R. and Ikeda, M.M. (1995). "A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys." Survey Methodology, 21-2, 147-157.

Hansen, Morris H., Hurwitz, William N., Madow, William G. (1993). *Sample Survey Methods and Theory*. Wiley, New York.

Kish, Leslie. (1965). *Survey Sampling*. Wiley, New York.

Kish, Leslie. (1992). "Weighting for Unequal Pi." Journal of Official Statistics, 8-2, 83-200.

Leaver, Sylvia G., Johnson, William H., Shoemaker, Owen J. and Benson, Thomas S. (1999). "Sample Redesign for the Introduction of the Telephone Point of Purchase Survey Frames in the Commodities and Services Component of the U.S. Consumer Price Index." Proceedings of the Section on Government Statistics and Section on Social Statistics, 292-297.

Rao, J.N.K., and Nigam, A.K. (1992). "'Optimal' Controlled Sampling: a Unified Approach." International Statistical Review, 60-1, 89-98.

Sitter, R.R., and Skinner, C.J. (1994). "Multi-way Stratification by Linear Programming." Survey Methodology, 20-1, 65-73.

Valliant, Richard, and Gentle, James E. (1997). "An Application of Mathematical Programming to Sample Allocation." Computational Statistics & Data Analysis, 25, 337-360.