THE USE OF MATHEMATICAL PROGRAMMING FOR DESIGNING DUAL FRAME SURVEYS

William F. McCarthy and David V. Bateman, U.S. Bureau of the Census
William F. McCarthy, FOB 3, Room 2068, Washington, DC 20233

KEY WORDS: sampling, dual frames

INTRODUCTION

This paper describes a methodology for determining the optimal allocation of sample units among dual frames. This methodology also allows the survey designer to conduct a post optimality analysis in order to study how various sample design and cost parameters affect the optimization of the dual frame sample allocation. In particular, a set of models which relate survey costs with statistical precision/accuracy, as well as other relevant constraints, are used. These models are used to create a mathematical program (optimization problem) such that the cost of conducting the survey is minimized subject to a prescribed level of statistical precision/accuracy or, conversely, the statistical precision/accuracy is maximized subject to an expected fixed cost for conducting the survey.

BACKGROUND

As was noted by Lund (1968), "A sampling frame or list is the keystone around which a sampling process is constructed". However, one often finds in actual practice that any one frame by itself may be inadequate to completely cover all units (households, persons, etc.) in the target population. Hartley (1962) found that by overlapping a list frame and an area frame, he could ensure a more complete coverage of the target population. This overlapping of two frames is referred to as a dual frame. In general, the list frame interviews are conducted via the mail, telephone, or personal visit (or some combination of the three). The area frame interviews are conducted via personal visit. Some examples of dual frames are:

A. DUAL FRAME, SINGLE MODE OF INTERVIEWING
   1. telephone interviewing (using random digit dialing) in conjunction with a telephone list.
   2. personal visits (using an address list) in conjunction with personal visits (using an area frame).

B. DUAL FRAME, MIXED MODE OF INTERVIEWING
   1. personal visits (using an address list) in conjunction with telephone (using random digit dialing).
   2. mail (using an address list) in conjunction with personal visits (using an area frame).

A number of researchers have looked at how one optimally allocates the sample among dual frames based on cost, variance, and bias estimates (Hartley, 1962; Lund, 1968; Casady, Snowden, and Sirken, 1981; Biemer, 1983; Lepkowski and Groves, 1986; and McCarthy and Bateman, 1988). As noted by Biemer (1983), "Although the concept of a dual frame survey is simple, the sample design issues can be very complex." He further states that the complexity of the various optimization formulae preclude most analytical investigations. This notion is corroborated by Arthanari and Dodge (1981): "The classical optimization methods based on differential calculus are too restrictive, and are either inapplicable or difficult to apply in many situations that arise in statistical work. This, together with the lack of suitable numerical algorithms for solving optimizing equations, has placed severe limitations on the choice of objective functions and constraints and led to the development and use of some inefficient statistical procedures." These latter researchers go on to say that mathematical programming has the potential for overcoming the problems associated with the classical optimization methods. In particular, they have used mathematical programming to optimally allocate sample sizes for stratified and multivariate stratified random sampling designs (single frames). In addition, Leaver, Weber, Cohen, and Archer (1987) used constrained integer nonlinear programming in determining optimum sample sizes for a single frame survey.

METHODOLOGY

A mathematical program is an optimization problem in which the objective and its constraints are given as mathematical functions and functional relationships. In general, they have the form

$$\text{optimize:} \quad f(\mathbf{X}) \qquad \text{[where optimize is maximize or minimize]}$$

$$\text{subject to:} \quad \begin{matrix} g_1(\mathbf{X}) \\ g_2(\mathbf{X}) \\ \cdots\cdots \\ g_m(\mathbf{X}) \end{matrix} \quad \begin{matrix} \leq \\ = \\ \\ \geq \end{matrix} \quad \begin{matrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{matrix}$$

$$\text{with } \mathbf{X} \geq \mathbf{0}.$$

Each of the m constraints involves one of the three signs $\leq$, =, $\geq$. In some cases, the additional restriction that $\mathbf{X}$ is integral is added. If $f(\mathbf{X})$ and each $g_i(\mathbf{X})$ (i=1,2,.....,m) are linear, then the mathematical program is linear. Any other mathematical program is considered nonlinear. The mathematical programs that are used in the design of dual frame surveys are nonlinear and usually have the restriction $\mathbf{X} \geq \mathbf{0}$ and integral.

For those readers seeking a review of the theory of mathematical programming (formulation and solution), they should refer to Himmelblau (1972), Phillips, Ravindran and Solberg (1976), and Arthanari and Dodge (1981). Those interested in how dual frame cost models are constructed should refer to Hartley (1962), Lund (1968), Casady, Snowden, and Sirken

(1981), Biemer (1983), Lepkowski and Groves (1986), McCarthy (1988A, 1988B), and McCarthy and Bateman (1988). All of the previously mentioned authors can also provide examples of dual frame statistical precision/accuracy models.

Mathematical programming has three major benefits for studying dual frame designs:

1. complex (both linear and nonlinear) cost models can be utilized;
2. the complex and sophisticated statistical precision/accuracy models associated with dual frame designs can be efficiently dealt with; and
3. post optimality analysis (sensitivity analysis) can be conducted to investigate how various sample design and cost parameters affect the optimization of the dual frame sample allocation.

An additional benefit is that there are software packages available for mathematical programming that all but eliminate the necessity to write customized computer programs. One such package is GINO (General Interactive Optimizer) produced by LINDO Systems Inc.

In general, a mathematical program dealing with a dual frame sample design would have the following form:

minimize: The overall total cost of conducting a dual frame survey
subject to: Statistical precision/accuracy model equal to some prescribed level; with all variables non-negative and integral.

Or alternatively, one can maximize the statistical precision/accuracy (i.e., minimize variance/total error) subject to a fixed overall total cost. It should be noted that other constraints such as time, staffing requirements, etc., can be utilized by such a program.

CONCLUSIONS

It is our belief that the use of mathematical programming will greatly help the survey designer in understanding and determining the feasible ranges of various dual frame design and cost parameters and their relative effects for dual frame optimization. In addition, the effects of sampling and non-sampling errors for dual frame designs can be systematically examined. For a comprehensive listing of the various sampling and non-sampling errors associated with dual frame designs the reader should refer to Biemer (1983) and Lepkowski and Groves (1986). Finally, the use of mathematical programming allows the survey designer the ability to not only deal effectively with complex and sophisticated cost and statistical precision/accuracy models but also with other relevant constraints such as interviewer workload distributions, staff distributions, the amount of time available for conducting the survey, etc.

BIBLIOGRAPHY

1. Arthanari, T.S. and Dodge, Y. (1981). Mathematical Programming in Statistics. John Wiley & Sons, N.Y.
2. Biemer, P.P. (1983). "Optimal Dual Frame Designs: Results of a Simulation Study", ASA Proceedings of Survey Research Methods, pp. 630-635.
3. Casady, R.J., Snowden, C.B., and Sirken, M.G. (1981). "A Study of Dual Frame Estimators for the National Health Interview Survey", ASA Proceedings of Survey Research Methods.
4. GINO (General Interactive Optimizer), developed by LINDO Systems, Inc., Chicago, IL 60614
5. Hartley, H.O. (1962). "Multiple Frame Surveys". ASA Proceedings of Social Statistics, pp. 203-206.
6. Himmelblau, D.M. (1972). Applied Nonlinear Programming. McGraw-Hill, N.Y.
7. Leaver, S.G., Weber, W.L., Cohen, M.P., and Archer, K.P. (1987). "Item - Outlet Sample Redesign, for the 1987 U.S. Consumer Price Index Revision". International Statistical Institute Proceedings. October.
8. Lepkowski, J.M. and Groves, R.M. (1986). "A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design", Journal of the American Statistical Association, 81 (396), pp. 930-937.
9. Lund, R.E. (1968). "Estimators in Multiple Frame Surveys", ASA Proceedings of Social Science.
10. McCarthy, W.F. (1988A). "OR/MS Applications in Computer Assisted Telephone Interviewing (CATI) Survey Research. Part I: Cost Modeling." The Institute of Management Sciences and Operations Research Society of America Joint National Meeting, April, Washington, D.C.
11. McCarthy, W.F. (1988B). "The Use of Cost Models and Mathematical Programming in Computer Assisted Telephone Interviewing (CATI)". For the 1989 National Conference Proceedings of the Institute of Cost Analysis, Washington, D.C.
12. McCarthy, W.F. and Bateman, D.V. (1988). "OR/MS Applications in Computer Assisted Telephone Interviewing (CATI) Survey Research. Part II: Optimal Allocation of Sample Sizes for Mixed Mode (Dual Frame) Survey Designs." The Institute of Management Sciences and Operations Research Society of America Joint National Meeting, April, Washington, D.C.
13. Phillips, D.T., Ravindran, A. and Solberg, J.J. (1976). Operations Research: Principles and Practice, Wiley & Son, N.Y.