SAMPLE REDESIGN FOR THE DRUG ABUSE WARNING NETWORK (DAWN)

James L. Green, Bob Baskin, and KC Lee, Westat
James L. Green, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Mathematical programming; Multi-variate sample design; Sample redesign

1. Introduction

The results of the sample redesign research conducted for the Substance Abuse and Mental Health Services Administration's (SAMHSA) Drug Abuse Warning Network (DAWN) are presented. The current DAWN sample was drawn in 1988 and research was conducted in 2000 and 2001 concerning a sample redesign, in addition to other research relating to case definition, data collection technology etc. A multi-variate sample design optimization problem is the core of the sample redesign, which also involved overlap control, sample maintenance, a phasing in strategy and two-phase sampling. This paper presents the research, approach and results of the redesign effort.

2. Background

DAWN uses a stratified, single-stage cluster sample of hospitals with 24-hour emergency departments. The AHA annual survey data base is used as the source for the sampling frame. Data are collected on all drug related emergency department visits. A drug related visit is called an episode. The particular drug(s) identified in the patient's chart for the ED visits are called mentions. Annual and semi-annual estimates of total episodes and mentions are required for the coterminous U.S. and for 21 specific Metropolitan Statistical Areas (MSAs). The national and MSA specific estimates are required for a number of drugs and with specified precision levels. These data are used to monitor changes in drug abuse patterns across time.

3. Current Design and Performance

The current sample design is a stratified, single stage sample of hospitals. A hospital is eligible for DAWN when it:

1. Is a general medical/surgical unit;
2. Has a 24-hour ED;
3. Is located within the coterminous U.S.;
4. Is a non-federal institution; and
5. Is a short-term stay institution.

The stratification is by MSA, size of unit (annual ED visits), central city status (located within the central city of the MSA or not) and type of services offered (presence or absence of an outpatient unit or an alcohol/chemical dependency unit). Units not located within the targeted 21 MSAs are assigned to a final primary stratum referred to as the national panel. The national panel is the key to producing a national estimate, given the targeted MSAs. Approximately 740 units have been sampled since 1988, of which 592 were eligible and 480 were responding in any given quarter of 1999. The sample was designed to yield 6 percent relative standard errors (RSEs) on estimates of total episodes for the national estimate and the three largest targeted MSAs (New York, Chicago and Los Angeles), 8 percent RSEs on estimates of total episodes otherwise. The sample is kept current through an annual sample maintenance process, which gives new or newly eligible units in the population a chance of selection. The current performance varies considerably by MSA and also by type of estimate (i.e., drug mentions) as shown in Table 1.

Table 1. Current performance–precision by MSA (RSEs)

Estimate	National	Min	Mean	Max
Total episodes	7.3%	0.4%	9.8%	21.3%
Cocaine mentions	9.0%	0.5%	13.0%	28.0%
Heroin mentions	13.5%	0.6%	12.7%	28.8%
Marijuana mentions	11.8%	0.4%	16.6%	41.7%

4. Redesign Objectives

SAMHSA had four objectives for the DAWN sample redesign, as follows:

1. Expand the coverage of the targeted MSAs;
2. Increase the precision of particular estimates;
3. Improve the stratification; and
4. Control the costs associated with the above.

SAMHSA desired expanded coverage of the targeted MSAs, as measured by the geographic dispersion of the sample, population coverage and increase in the number of targeted MSAs. The agency also desired increased precision for particular MSAs, and increased precision for a number of estimates in addition to total episodes. Improved stratification was considered part of the solution to meeting these objectives, while it was also realized that costs would have to be controlled while trying to meet these objectives.

5. Overall Sample Design Optimization Process

5.1 Sample Design and Sample Size Estimation

Sample design and sample size estimation requires the following four general classes of parameters:

1. Estimates (e.g., total episodes, cocaine mentions etc.);
2. Level (e.g., national, region, MSA);
3. Precision requirements (e.g., RSEs); and
4. Costs (e.g., direct reimbursement or general data collection).

A common reference method was required in order to rapidly generate, solve, and review sample size estimation problems in an iterative process. We referred to a given combination of the above parameter classes as a sample design scenario, which was the basis for communication with the agency.

5.2 Overall Process

The sample design optimization process (see Figure 1) consisted of the following steps:

1. Model and calculate population values;
2. Generate sample design scenarios;
3. Solve the mathematical programming problem;
4. Obtain the results and design; and
5. Review with SAMHSA and iterate.

5.3 Optimization Problem

We used a mathematical programming approach to solving our multi-variate sample design problem. Mathematical programming has been used to solve a variety of sample design and allocation problems (see Arthanari and Dodge, 1993 (Chapter 5); and Green, 2000 for a summary), and has been used to solve the multi-variate sample design problem extensively (see Bethel, 1989; Leaver et al, 1999; and Valliant and Gentle, 1997). We used SAS PROC NLP, within the SAS OR routine library. We solved both a sample size minimization and a cost minimization version of our problem. The sample size minimization version of the problem can be expressed as follows:

Minimize: $$T = \sum_{i=1}^{L} \sum_{j=1}^{H_i} n_{ij}$$

Subject to: $$2 \le n_{ij} \le N_{ij}$$
$$RSE_i(x_k) \le T_i(x_k)$$

where i = the primary stratum index (e.g., MSA)

j = the secondary stratum index (e.g., ownership *X* size)

L = the number of primary strata (number of specifically targeted MSAs + 1 for the national panel stratum)

H_i = the number of secondary strata in primary stratum i

n_{ij} = the sample size in stratum ij (these are the decision variables)

N_{ij} = the population size in stratum ij

$RSE_i(x_k)$ = the relative standard error of estimate k in stratum i

$T_i(x_k)$ = the target relative standard error of estimate k in stratum i

and the relative standard error is expressed as follows:

$$RSE_i(x_k) = \frac{\sqrt{\sum_{j=1}^{H_i} \frac{W_j^2 S_j^2 (1 - f_j)}{n_j}}}{\bar{x}_{ik}}$$ for MSA estimates and

$$RSE_N(x_k) = \frac{\sqrt{\sum_{i=1}^{L} \sum_{j=1}^{H_i} \frac{W_{ij}^2 S_{ij}^2 (1 - f_{ij})}{n_{ij}}}}{\bar{x}_{Nk}}$$ for the national estimates.

A cost minimization version of the problem can be expressed as follows:

Minimize: $$T = \sum_{i=1}^{L} \sum_{j=1}^{H_i} c_{ij} n_{ij}$$

Subject to: $$2 \le n_{ij} \le N_{ij}$$
$$RSE_i(x_k) \le T_i(x_k)$$

where: $$C_{ij} = F(\bar{B}_{0ij} + \bar{B}_{1ij}\bar{x}_{ij} + \bar{B}_{2ij}\bar{y}_{ij})Z_{ij}$$

F = 0.90 (this was to provide room for negotiation with the unit)

Z_{ij} = a cost of living adjustment at the MSA level

Since the coefficients for the current direct reimbursement model were not constant within the strata proposed for the new sample design, we

calculated an average annual direct reimbursement cost for each unit within a particular stratum as follows:

where $\bar{x}_{ij}$ = the mean annual ED visits within stratum *ij*

$\bar{y}_{ij}$ = the mean annual total episodes within stratum *ij*

$\bar{B}_{0ij}$ = the mean minimum compensation level within stratum *ij*

$\bar{B}_{1ij}$ = the mean ED visits coefficient from the current cost model within stratum *ij*

$\bar{B}_{2ij}$ = the mean total episodes coefficient from the current cost model within stratum *ij*

6. Recommended Design and Performance

To ensure adequate sample sizes for the precision required within targeted MSAs, we chose MSA as the primary stratification variable. Within MSAs, secondary stratification was by ownership (public, non-public) and size (annual ED visits). Up to 4 size strata were used within each ownership category. These secondary stratification variables were determined to be the most effective in partitioning the variance on annual ED visits.

After running numerous scenarios through the process described in Section 5.2, the review and iteration allowed SAMHSA to identify the analytic objectives and design parameters that seemed to be within the projected resources. This included finalizing the list of targeted MSAs, which ultimately included the top 5 MSAs (in terms of population) within each of the 9 Census divisions, plus any of the current 21 MSAs otherwise not included. The final sample design parameters consisted of the following:

1. Estimates (total episodes, cocaine mentions, heroin mentions, marijuana mentions);
2. Level (MSAs, national); and
3. Precision requirements (10 percent RSEs for MSA estimates, 15 percent RSEs for the national estimates).

The expected performance of the new design is presented in Table 2.

The difference between the sample size minimization and cost minimization results, in terms of sample size and estimated direct reimbursement costs, are presented in Table 3. The differences between the two approaches were considered trivial, and the results from the sample size minimization approach were accepted as the recommended design.

Table 2. Expected new design performance – precision by MSA (RSEs)

Estimate	National	Min	Mean	Max
Total episodes	9.0%	0.0%	3.9%	10.0%
Cocaine	13.0%	0.0%	6.3%	10.0%
Heroin	15.0%	0.0%	7.2%	10.0%
Marijuana	15.0%	0.0%	7.8%	10.0%

Table 3. Comparison of sample size and cost minimization results

Optimization problem	Required responding sample size	Estimated direct reimbursement
Minimize sample size	949	$3,361,490.41
Minimize cost	961	$3,334,997.53

7. Other Design Features

Other recommended sample design features included overlap control, sample maintenance, a phasing in strategy and two-phase sampling.

7.1 Overlap Control

Overlap control was recommend to be considered as a recruitment cost reduction technique. Research indicated that the information required to do overlap control was available from the previous design. However, overlap control had been previously used in DAWN in the 1988 sample design. This meant that the 1988 selection could not be treated as an independent selection within stratum. There was also change in the definition of the strata which makes the overlap more complicated. The overlap control was investigated by calculating the expected overlap under an independent selection and then calculating the expected overlap using an overlap control method. The expected improvement in overlap between an independent selection and a selection using overlap control varied considerably across the MSAs included in both the old and new designs.

The methodology used for the overlap control selection is an extension of the method of Keyfitz. The method is discussed in Chowdhury, Chu, and Kaufman (2000). Westat has written proprietary software to perform this overlap control in the form of a SAS

macro. This allows for efficient investigation of the overlap problem. This particular method of overlap control is optimal but not exact. It is not exact because the sample size can vary in the sense that the input sample size is achieved on average but any given solution of the overlap control problem may not produce exactly the sample size that was input. However, given the sample size that was produced as the solution, the solution is optimal in the sense that the expected overlap cannot be exceeded by any other solution. The fact that the solution is approximate was not deemed to be a problem since the input sample sizes were solutions to an optimization problem and these solutions varied with the specification of that problem. There was no reason from a optimization point of view or from procedural point of view to achieve the exact sample sizes.

As an example of the overlap problem the Boston MSA is a good case. There are eight new sample strata in Boston. Of the eight strata, two strata are empty, three strata are certainties, and in one stratum there are no previously selected units. Thus, in six of the eight strata the overlap control has no effect. This is somewhat typical of the 21 MSAs currently in the DAWN sample. In the remaining two strata there are 17 and 18 units of which 9 and 14 are to be chosen, respectively. Sampling at random would produce overlap in these two strata of 3.71 and 7.02 whereas the expected overlap from the controlled procedure would be 5.59 and 8.56 respectively. From the other six strata there are 21 guaranteed overlaps. Thus, the total expected overlap for independent sampling would be 31.73 units versus 35.15 units for controlled overlap. In some of the 21 MSAs the overlap control would not be used because all units were selected in the 1988 design but in other MSAs the overlap methodology could produce an improved expected overlap.

7.2 Sample Maintenance

Sample maintenance was also recommended as had been previously used, in order to allow new or newly eligible units a chance of selection and keep the DAWN sample up to date. We also recommended consideration of a panel rotation scheme, which may allow sampled units that have become ineligible to be removed and replaced periodically. The latter may be important in particular targeted MSAs which experience considerable change in the health care system.

7.3 Phasing in Strategy

A phasing in strategy was required, as SAMHSA did not envision having the resources to expand from 21 to 48 MSAs immediately. Two options were proposed, as follows:

1. Expansion by sample size across MSAs; and
2. Expansion by MSA.

Expansion by sample size across MSAs would bring all 48 targeted MSAs into the new sample from the start, and gradually expand to the sample sizes required for the full desired precision over time. The phasing in period may be something like 4 years. Expansion by MSA would bring some number of the 48 targeted MSAs into the new sample each year, but at the sample sizes required for the full desired precision. The relative advantages and disadvantages to the two approaches thus relate to the coverage and precision objectives, the value given to each, and SAMHSA's resultant priorities. Both approaches also have implications for the national panel and the resulting precision of the national estimate. There would be clever ways to structure the national panel, given the ultimate distribution of the sample, under either approach. The precision of the national estimate could be protected at any given moment, at the expense of deselecting some national panel units after full expansion.

7.4 Two-Phase Sampling

We researched the possibility of two-phase sampling using the historical data available from DAWN. Our suggestion was to consider using the first phase to obtain better measure of size information (total episodes, cocaine mentions etc.), from which the second phase sample would be drawn. Currently only annual ED visits is available from the AHA-based frame. The feasibility of two-phase sampling depends on the following:

1. The predictive ability of the 1st phase data; and
2. The relative 1st and 2nd phase data collection costs.

We measured the predictive ability of a few months of data vis-à-vis annual reported counts of events. The results were encouraging. Depending upon the estimate of interest, simple regression models yielded r-square measures of 0.73 to 0.99 (see Table 4), indicating that a large amount of variance in the annual measure was explained by the few months of data.

The regression model used was as follows:

$$\hat{Y}_{ijk} = B_{0jl} + B_{1jl} X_{ijl},$$

where

$\hat{Y}_{ijk}$ = the total annual estimated count of episodes for unit i, estimate j, 12 month period k

B_{0jl} = the intercept term for estimate j, 2 month period l

B_{1jl} = the coefficient term for estimate j, 2 month period l

X_{ijl} = the reported count of episodes for unit i, estimate j, 2 month period l

Table 4. Minimum, mean and maximum (across eleven two month models) r-square measure for two-phase models

Estimate	Min.	Mean	Max.
ED Visits	0.9684	0.9792	0.9862
Total episodes	0.9601	0.9696	0.9768
Cocaine episodes	0.9622	0.9710	0.9860
Heroin episodes	0.9337	0.9459	0.9679
Marijuana episodes	0.9263	0.9544	0.9817
Methamphetamine episodes	0.7258	0.8343	0.9341

The relative 1st and 2nd phase data collection costs will determine if the 1st phase predictive ability can be utilized in a cost-effective way. This will depend on the cost structure of the new implementation and should be evaluated further at that time.

8. References

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

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

Chowdhury, S., Chu, A., and Kaufman, S. (2000). Minimizing Overlap in NCES Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Green, J.L. (2000). Mathematical Programming for Sample Design and Allocation Problems. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 688-692.

Leaver, S.G., Johnson, W.H., Shoemaker, O.J., and Benson, T.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.

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

Figure 1. DAWN Sample Redesign – Sample Design Scenarios Evaluation System