SAMPLE ALLOCATION FOR THE STATUS OF THE ARMED FORCES SURVEYS

R. E. Mason, S. C. Wheeless, B. J. George, and J. A. Dever, Research Triangle Institute
R. A. Riemer and T. W. Elig, Defense Manpower Data Center
R. E. Mason, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, NC 27709

Key Words: Optimization, Sample Allocation

1. Introduction

The 1995 Status of the Armed Forces Surveys (SAFS) deal with gender, racial, and ethnic issues in the United States military establishment. A total of four surveys are involved. The examples used in this paper are drawn from one of the four surveys, known as the Form B survey, which deals with gender issues.

Each survey includes members of the four Armed Services, the Coast Guard, National Guard, and Reserves worldwide. Data collection is by mail. Sample individuals initially receive an introductory letter that explains the survey and solicits cooperation. The letter is followed by a package containing the questionnaire and instructions for completing and returning the information. The package is followed by a second letter thanking the individual for having returned the questionnaire or otherwise asking for its return. After a specified time has elapsed a second package containing the questionnaire and instructions is mailed to nonrespondents.

An unusual feature of these surveys is the large amount of information that is available for design purposes about the individuals that comprise the population. Not only are the demographics of individuals known in some detail, but so also are their occupational specialties, their work locations and settings, and their positions within the total organizational structure. This wealth of concomitant information is used to control the distribution of the sample for the purpose of providing predetermined levels of precision for estimates of parameters that describe key reporting domains.

The information is used to construct strata and to determine the sizes of the key reporting domains within each of the defined strata. Given the stratum sizes and their composition, variance constraints are placed on parameter estimates describing domains defined within one or more strata and overall. Equations are developed that describe the variances of the estimates and the variable survey costs in terms of the salient features of the design, which are constants in the equations, and the sample sizes to be allocated as specified by the design structure, which are the unknowns in the equations. The equations are solved simultaneously subject to the variance constraints to yield that allocation of the total sample that jointly satisfies the imposed variance constraints for the least cost.

This method for determining a sample allocation was first developed by J. R. Chromy for use in a medical provider record check survey conducted by the Research Triangle Institute in the late 1970s (Folsom et al. (1979)). The procedure is described in Chromy (1987).

The variance equations, of course, require knowledge of the relevant population variances. In practice the population variances are likely to be unknown, at least in advance of the survey, which is the case for these surveys. We have, as a consequence, defined the parameters of interest to be population proportions such that the (binomial) population variances are coincidentally specified with specifications for the values of the proportions. That is, the parameters of interest for determining the sample allocation are the relative sizes of specified key domains. The convention introduces some generality and provides a useful surrogate for other parameters. Certainly parameters describing other domain characteristics are unlikely to be reliably estimated if the domain sizes themselves cannot be. This choice of parameters is not restrictive if the requisite population variances are known.

2. Sampling Design

A stratified random sampling design is used for the SAFS. Sample individuals are selected with equal conditional probabilities given the stratum and without replacement.

The dimensions of stratification are shown in Table 1 along with the maximum number of levels in each dimension. The dimension labeled as Unknown contains all individuals for which at least one of the variable values needed to identify the appropriate level of stratification is missing from the source files used to construct the sampling frame. The stratum sizes resulting from forming all possible crosses of levels within dimensions were computed and compared with the minimum stratum size consistent with a proportional allocation of a total sample size of 40,000.

Table 1. Dimensions And Levels Of Stratification

Dimension	Levels
Service	Army
	Navy
	Marine Corps
	Air Force
	Coast Guard
	Reserves and National Guard (AGR/TARS)
Location	Continental United States (CONUS)
	Outside Continental United States (OCONUS)
Pay Grade	Enlisted Grades E1-E4
	Enlisted Grades E5-E9
	Company Grade and Warrant Officers
	Field Grade Officers
Gender	male
	female
Race/Ethnicity	non-Hispanic White
	non-Hispanic Black
	Hispanic any race
	Other
Unknown	

Stratum cells smaller than the minimum were identified as candidates for collapsing into other cells

In undertaking the collapsing, the dimensions of stratification were considered to be nested in the order in which they are presented in Table 1. First, racial categories for females overall were collapsed into two levels, non-Hispanic White and Other, except for female Marine Corps officers stationed overseas for whom no racial categories were defined. Second, locations were collapsed within the Coast Guard and within the National Guard and Reserves combination. A total of 180 strata were constructed.

Key reporting domains at the level of the overall population were defined using the same variables and variable values as were used for stratification with one addition. The addition involved occupations, with domains defined by the representation of women in an occupation. Occupation specialties in the military are different for officers and enlisted personnel. In each case the relevant list of occupations was divided into quartiles based on the proportion of women. Within the first quartile, which might be described as the most extremely male dominated occupations, four domains were defined to further identify those occupations with the very lowest representation of women. Otherwise the domains were defined by the quartiles of the distribution, making a total of seven occupational domains.

The domain sizes used to allocate the sample are the gender specific proportions of persons reporting at least one of the behaviors that define unwanted sexual attention. Domains defined at the level of the overall population are termed main effect domains in what follows. First order interactions are defined by crossing pairs of main effect domains, for example, gender by race. Higher order interactions are similarly defined. In addition to being important in their own right, variance constraints imposed on main effect domains act to control unequal weighting effects induced by the total pattern of imposed constraints, particularly those imposed on the higher order interactions (i.e., smaller domains).

The number of main effect, first and second order interaction domains used to allocate the sample together with their associated variance constraints are shown in Table 2. The precision requirements cited in Table 2 are confidence interval half-widths.

3. Sample Allocation

The variance constraints take the form,

$$v_d(n_s) \le K_d \ , \qquad d = 1,2,\ldots,D$$

where $v_d(n_s)$ is the variance function for the d-th parameter estimate and K_d is the constraint imposed by the investigator. The form of the variance function is, of course, specified by the design. The notation is intended to suggest that, regardless of its form, the variance is a function of unknown sample sizes, n_s.

Table 2. Variance Constraints

Domain Description	Number of Domains	Precision Requirements
Gender	2	0.02
Location	2	0.03
Service [1]	6	0.05
Gender by Occupation	14	0.08
Gender by Race	8	0.05
Gender by Location	4	0.03
Gender by Service [1]	12	0.05
Females by Pay Grade Group [3]	6	0.03
Females, Enlisted by Service [1]	6	0.05
Females, Commissioned and Warrant Officers by Service [1]	6	0.05
Females, E1-E3, by Active Duty Service [2]	5	0.05
Females, E4, by Active Duty Service [2]	5	0.10
Females, E5-E6 by Active Duty Service [2]	5	0.10
Females, E7-E9 by Active Duty Service [2]	5	0.10
Females, Company Grade Officers by Active Duty Service [2]	5	0.10
Females, Field Grade Officers by Active Duty Service [2]	5	0.10
Males by Pay Grade Group [3]	6	0.05
Males, Enlisted by Service [1]	6	0.06
Males, Commissioned and Warrant Officers by Service [1]	6	0.06
Males, E1-E3 by Active Duty Service [2]	5	0.06
Males, E4-E9 by Active Duty Service [2]	5	0.06
Total	124	

[1] Army, Navy, Marine Corps, Air Force, Coast Guard, National Guard and Reserves

[2] Army, Navy, Marine Corps, Air Force, Coast Guard

[3] E1-E3, E4, E5-E6, E7-E9, Company Grade Officers, Field Grade Officers

In addition to the variance function, a cost function $c(n_s)$ is developed to describe the total variable cost of the survey in terms of the same unknown sample sizes. Variable costs may, in general, be both domain and stratum specific. The cost modeling exercise is, therefore, to develop equations that describe the domain and stratum costs as appropriate and then combine them in the proper proportions to obtain the overall cost.

Given the cost and variance functions, interest lies in determining the values ${}^*n_s$ that minimize the objective function,

$$o(n_s,\lambda_d) = c(n_s) + \sum_d \lambda_d (v_d(n_s) - K_d),$$

where the λ_d are generalized Lagrange multipliers, one for each of the variance constraints imposed. Taking derivatives of the objective function yields equations of the form,

$$-\frac{\partial(c(n_s))}{\partial(n_s)} = \sum_d \lambda_d \frac{\partial(v_d(n_s))}{\partial(n_s)}. \qquad [1]$$

If the variance constraints hold, then at ${}^*n_s$ there must exist values of the Lagrange multipliers ${}^*\lambda_d$ such that equation [1] evaluated at ${}^*n_s$ is true and additionally,

$$v_d({}^*n_s) \le K_d\ , \qquad [2]$$

$${}^*\lambda_d \ge 0\ , \qquad [3]$$

$${}^*\lambda_d (v_d({}^*n_s) - K_d) = 0\ . \qquad [4]$$

Equations [1] through [4], with ${}^*n_s$ substituted in equation [1], are the Karush-Kuhn-Tucker necessary conditions (Kuhn and Tucker (1951)). Sufficiency is

argued on the basis that the cost function $c(n_s)$ is a convex function and the constraints $K_d - v_d(n_s)$ are concave functions (see, for example, Hillier and Lieberman (1974), pages 722 through 725).

3.1 Variance Model

Define the indicator variables

$\delta_{d,h,i} = 1$, if the i-th individual in the h-th stratum belongs to the d-th domain,
$= 0$, otherwise,

$\delta_{h,i} = 1$, if the i-th individual in the h-th stratum reports having experienced at least one of the behaviors defining unwanted sexual attention,
$= 0$, otherwise.

Then the total members of the domain who report having experienced at least one of the behaviors is the quantity

$$N_d P_d = \sum_h \sum_{i=1}^{N_h} \delta_{d,h,i} \delta_{h,i}$$

where $i = 1, 2, \ldots, N_h$ identifies the individuals classified into the h-th stratum. The relative domain size is the population proportion

$$P_d = \sum_h \frac{N_h}{N_d} P_{d,h}$$

where

$$P_{d,h} = \frac{1}{N_h} \sum_{i=1}^{N_h} \delta_{h,i} \delta_{d,h,i} .$$

Denote the sample estimate of the proportion by

$$\hat{P}_d = \sum_h \frac{N_h}{N_d} \hat{P}_{d,h} ,$$

with variance

$$v(n_s) = Var\{\hat{P}_d\} = \sum_h \left(\frac{N_h}{N_d}\right)^2 Var\{\hat{P}_{d,h}\}$$

where, if the stratum-level samples are selected with equal probability and without replacement,

$$Var\{\hat{P}_{d,h}\} = \left(\frac{N_h - n_h}{N_h - 1}\right)\left(\frac{P_{d,h}(1-P_{d,h})}{n_h}\right). \qquad [5]$$

At the level of an individual domain, the variance constraints in Table 2 are of the form

$$Var\{\hat{P}_d\} \le K_d = \left(\frac{CI\{\hat{P}_d\}}{1.96}\right)^2$$

where $CI\{\hat{P}_d\}$ are the confidence interval half-widths reported in Table 2.

3.2 Cost Model

A candidate list of activities to be potentially included in a cost model consists of the following items:

- sampling frame construction
- sample selection
- instrument development
- data collection
- data editing
- data processing
- data analysis and reporting

For the SAFS, with a single stage of sampling, variable survey costs are largely if not quite completely defined by the data collection, data editing, and data processing activities. Cost coefficients can be developed for these activities in terms of the per unit cost of packages sent out on the first and second mailings, C_1 and C_2, and on the per unit costs of packages that are returned, C_3.

The cost model takes the form,

$$c(n_s) = \sum_h n_h \bar{C}_h$$

where, denoting the response rates to the first and second mailing by R_1 and R_2 respectively,

$$\bar{C}_h = \frac{C_{1,h} + (1 - R_{1,h}) C_{2,h} + (R_{1,h} + R_{2,h}) C_{3,h}}{R_{1,h} + R_{2,h}} .$$

The h-subscripts allow the cost coefficients and response rates to be different in different strata if

appropriate. Military postal services are used for these surveys such that the cost coefficients are the same in all strata. However response rates were allowed to be different according to Service, pay grade, gender, race and ethnicity based on current experience with related surveys.

3.3 Allocation Solutions

Taking derivatives of the objective function with respect to the stratum-level sample sizes, equating to zero, and solving for the values n_h yields solutions of the form,

$$n_h = \sqrt{\frac{\sum_d \lambda_d \left(\frac{N_h}{N_d}\right)^2 \left(\frac{N_h}{N_h - 1}\right) P_{d,h}(1 - P_{d,h})}{\bar{C}_h}}.$$

The solutions ${}^*n_h$ and ${}^*\lambda_d$ are found numerically. If to start the numerical procedure the initial values of the Lagrange multipliers are set to

$$\sqrt{{}^0\lambda_d} = \frac{\sum_h \left(\frac{N_h}{N_d}\right)\left(\sqrt{P_{d,h}(1 - P_{d,h})}\right)\left(\sqrt{\bar{C}_h}\right)}{K_d},$$

then a comparison of the initial values ${}^0\lambda_d$ and the final values ${}^*\lambda_d$ will identify those variance constraints that exert the major influence in determining the sample allocation and, by implication, the cost of the survey. In general the initial values chosen for this purpose are those values of the Lagrange multipliers that satisfy the constraints individually. Then final values that are closest to these initial values identify those constraints that are the most important in determining the allocation. A small relaxation of the identified constraints can yield important reductions in the variable cost of the survey should the initially imposed constraints prove unaffordable. Constraints that are satisfied coincidentally with the imposition of other constraints will have final Lagrange multiplier values of zero.

4. Results

The variance constraints listed in Table 2 were determined over several iterations. Our initial specifications of the constraints proved too restrictive to be practical. At each iteration, those constraints that were the major determinants of the allocation solutions were identified and progressively relaxed until a set of constraints were developed that provided both an informative and an affordable study. Given the specifications in Table 2, the ten constraints that were the major determinants of the final allocation solutions are listed in order in Table 3.

Note that all of the constraints in Table 3 are second order interactions. This result is not surprising in that such constraints involve quite fine subdivisions of the total population. The first order interaction that is the most important in determining the sample allocation is that imposed on female field grade officers, with a Lagrange multiplier ratio of 0.8820. By comparison, all of the main effect constraints have ratios that are essentially zero, indicating that the constraints were coincidentally satisfied with the imposition of the other constraints.

Because the imposed constraints are inequality constraints, the average performance of the sample in general tends to be better, that is, tends to have smaller variances, than is suggested by the constraints themselves. Table 4 reports the range of confidence interval half-widths computed using the allocation solutions for comparison with the requirements listed in Table 2.

Shown also in Table 4 are the ranges of the design effects for the domain estimates. The major component of the design effect is, of course, the unequal weighting effect associated with the disproportionate sample allocation.

Design effects judged to be excessively large provide some guidance for modifying either the design strata or the domain constraints or both. For example, the Service associated with the design effect of 7.51 in Table 4 is the Coast Guard. The efficiency of the design for this main effect constraint could perhaps be improved by removing racially defined strata, as was done for the female Marine Corps stationed overseas, and collapsing pay grade strata. Alternately, or in addition, the variance constraints imposed on the Coast Guard higher order interaction domains could be relaxed.

5. References

Chromy, J. R., 1987, Design Optimization With Multiple Objectives, Proceedings of the Section on Survey Research Methods, American Statistical Association, p. 194-199.

Table 3. List of Ten Most Restrictive Constraints

Domain Description	$\frac{^*\lambda_d}{^o\lambda_d}$
Females, Field Grade Officers, Coast Guard	0.9964
Females, E7-E9, Coast Guard	0.9955
Females, E1-E3, Coast Guard	0.9932
Males, E1-E3, Coast Guard	0.9883
Males, Officers, Coast Guard	0.9852
Females, Field Grade Officers, Marine Corps	0.9795
Males, E1-E3, Air Force	0.9561
Females, E1-E3, Marine Corps	0.9401
Males, Officers, AGR/TAR	0.9191
Males, Officers, Marine Corps	0.9126

Table 4. Variances and Design Effects

Domain Description	Precision	Design Effects
Gender	0.009 to 0.014	1.34 to 2.01
Location	0.014 to 0.027	3.82 to 5.54
Service	0.022 to 0.042	3.96 to 7.51
Gender by Occupation	0.021 to 0.080	1.66 to 4.07
Gender by Race	0.012 to 0.050	1.13 to 2.71
Gender by Location	0.014 to 0.027	1.10 to 2.11
Gender by Service	0.016 to 0.050	1.06 to 1.71
Females by Pay Grade Group	0.012 to 0.030	1.48 to 2.02
Females, Enlisted by Service	0.019 to 0.029	1.00 to 1.49
Females, Commissioned and Warrant Officers by Service	0.020 to 0.040	1.00 to 1.09
Females, E1-E3 by Active Duty Service	0.046 to 0.050	1.31 to 1.63
Females, E4 by Active Duty Service	0.048 to 0.077	1.63 to 1.92
Females, E5-E6 by Active Duty Service	0.023 to 0.032	1.17 to 1.23
Females, E7-E9 by Active Duty Service	0.050 to 0.086	1.78 to 1.88
Females, Company Grade Officers by Active Duty Service	0.027 to 0.037	1.30 to 1.44
Females, Field Grade Officers by Active Duty Service	0.046 to 0.087	1.67 to 1.78
Males by Pay Grade Group	0.029 to 0.050	1.50 to 1.80
Males, Enlisted by Service	0.029 to 0.060	1.01 to 1.11
Males, Commissioned and Warrant Officers by Service	0.053 to 0.059	1.00 to 1.01
Males, E1-E3 by Active Duty Service	0.059 to 0.060	1.12 to 1.24
Males, E4-E9 by Active Duty Service	0.036 to 0.060	1.11 to 1.90

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

Hillier, F. S., and G. J. Lieberman, 1974, Operations Research, Second Edition, Holden-Day, Inc., San Francisco CA, 800p.

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