An Optimal Allocation Method for Two-Stage Sampling Designs with Stratification at the Second Stage

Jill A. Dever, Jun Liu, Vincent G. Iannacchione, and Douglas E. Kendrick, Research Triangle Institute
Jun Liu, Research Triangle Institute, 3040 Cornwallis Rd, PO Box 12194, Research Triangle Park, NC 27709-2194

Keywords: Stratified Two-stage PPS Design, Variance Components, Cost/Variance Optimal Sample Allocation, Design Effects.

1. INTRODUCTION

Sampling designs for large-scale surveys often specify two stages of selection: clusters or primary sampling units (PSUs) at the first stage; and, second-stage units (SSUs) that are classified into strata within PSUs. The PSUs are selected with probabilities proportional to size and without replacement from a set of first-stage strata. Once the design is chosen, one must. We described a non-linear optimization procedure for determining the number of PSUs and SSUs to select given a set of constraints in a paper presented at the Joint Statistical Meeting in Dallas, TX (Liu, Iannacchione, Kavee, 1998). Our procedure accounts for stratification at the first and second stages and uses design-consistent variance components to improve the optimization procedure. Many other design optimizations only account for stratification at the first stage. In this paper, we demonstrate the utility of our proposed optimization scheme by comparing recent results with those from the prior round of a survey.

2. APPLICATION

The sample design for the is a stratified two-stage design as described above. We implemented our proposed optimization procedure for the *Department of Defense Survey of Health Related Behaviors Among Military Personnel* (*DoD Survey*) (Iannacchione, et al 1998). The stratified two-stage design for this study specified that military installations (PSUs) be selected from eight first-stage strata – branch of the military (Army, Navy, Marine Corps, and Air Force) by PSU location either inside (CONUS) or outside (OCONUS) the continental United States. Within PSUs, stratified random samples of active-duty military personnel (SSUs) were selected independently from 12 second-stage strata within the selected PSUs – gender by six military rank categories.

Our task was to determine the number of PSUs and SSUs for the study that would satisfy the analytical requirements and the fiscal constraints imposed on the survey. We set up a nonlinear optimization problem using the Kuhn-Tucker conditions (Chong and Zak, 1996) to search for the optimal sample size and allocation between the strata. This iterative process minimizes a cost function while meeting or exceeding a set of precision constraints. The cost function was constructed in terms of the first- and second-stage sample sizes, n_h and m_{hk}, respectively:

$$C = C_0 + \sum_h \left(c_{1h} n_h + \sum_k c_{2hk} m_{hk} \right)$$

where C_o is the fixed cost and is assumed zero in the optimization. Parameters c_{1k} and c_{2hk} are the variable cost associated with adding an additional PSU and SSU, respectively, to the survey. The cost model was used to account for the increased expense of sampling OCONUS installations relative to the CONUS PSUs, and Navy afloat units relative to the ashore units..

In the next section of the paper, we discuss the remaining piece of the optimization procedure - formulas for the design-consistent variance components. We placed constraints on the variance of a set of outcomes as well as constraints on the overall sample sizes for the *1998 DoD Survey*. Examples of the constraints are provided later in the paper.

3. VARIANCE COMPONENTS

As discussed earlier, we focus on stratified two-stage sample designs where the second-stage strata are nested within the PSUs. Because the sampling

method of the second-stage units does not affect the variance formula, we will present the result where simple random sampling is used within the second-stage strata.

For the discussion of variance components used for the *1998 DoD Survey*, let us first define the design and the variable subscripts used in subsequent formulas. The first-stage sampling frame was stratified into eight first-stage strata, indexed by *h*. The SSUs were stratified into 12 second-stage strata, indexed by *k*. The PSUs (indexed by *i*) were selected using a without-replacement PPS procedure. A random sample of SSUs was selected independently within each second-stage stratum.

Standard textbooks demonstrate how the overall variance can be decomposed into first-stage and second-stage components, or "between-PSU" and "within-PSU" components.

$$\sigma^2 = \sigma^2_{PSU} + \sigma^2_{SSU}$$

Prevalence estimates ($\hat{p}$) from the questionnaire were used to estimate the population variance. These estimates included alcohol use, alcohol abuse, illicit drug use, cigarette smoking, and smokeless-tobacco use. We calculated the prevalence estimates for a set of domains - Service, rank, gender, and branch by gender. The median prevalence estimate was used over a single value to produce a more stable estimate. The domains of interest are indexed by *d* and the population size within domain *d* is M_d. When M_d is known for the d^{th} domain, the domain proportion (p_d) can be estimated using the following linear estimator,

$$\hat{p}_d = \frac{1}{M_d} \sum_h \hat{y}_{dh}$$

where,

$$\hat{y}_{dh} = \frac{1}{n_h} \sum_i \frac{\hat{y}_{dhi}}{z_{hi}}$$

is the Horvitz-Thompson (1952) estimator for stratum *h*. As described in our 1998 paper, the following is the decomposition for the variance of $\hat{p}_d$. We leave the details to our initial paper and provide a overview below.

$$Var(\hat{p}_d) = Var_{PSU}(\hat{p}_d) + Var_{SSU}(\hat{p}_d)$$
$$= \sum_h \left(\frac{\hat{\sigma}^2_{PSU,dh}}{n_h} + \frac{\hat{\sigma}^2_{SSU,dh}}{n_h \overline{m}_h} \right)$$

Continuing, we can rewrite the formulas presented in our first paper into the following form for the first-stage variance component

$$\hat{\sigma}^2_{PSU,dh} = \frac{m_h}{M_d^2 (n_h - 1)} \sum_i^{n_h} \left(\frac{\hat{y}_{dhi}}{C_{hi}} - \frac{\hat{Y}_{dh}}{m_h} \right)^2 - \sum_{k \in D_d} \sum_i^{n_h} \left(\frac{M_{hk} M_{hik}}{C_{hi}} \times \frac{(1 - f_{hik}) s^2_{2hik}}{m_{hk}} \right)$$

and the second-stage variance component

$$\hat{\sigma}^2_{SSU,dh} = \frac{m_h}{n_h M_d^2} \sum_{k \in D_d} \sum_i^{n_h} \left(\frac{m_k (1 - f_{hik}) s^2_{shik}}{C_{hi} m_{hk}} \right)$$

where,

$s^2_{shik} = \hat{p}_{hik}(1 - \hat{p}_{hik})$;

$\hat{y}_{dhi}$ is the estimated total within the i^{th} PSU in the h^{th} first-stage stratum;

m_{hk} is the number of sampled individuals in the k^{th} second-stage stratum within the h^{th} first-stage stratum;

M_{hik} is the total number of individuals in the population within the k^{th} second-stage stratum within the i^{th} PSU of the h^{th} first-stage stratum; and

s^2_{2hik} is the estimated variance within the k^{th} second-stage stratum within the i^{th} PSU of the h^{th} first-stage stratum.

A point of interest is the composite size measure (C_{hi}) or CSM displayed in both variance component formulas above. CSMs were first proposed by Folsom, et al. (1987) for use as a tool to achieve self-weighting designs when oversampling second-stage domains. The CSM was used in the *1998 DoD Survey* to select military installations (PSU) with a higher concentration of females to increase their representation in the sample. This concentration varies across Service and across installations within

Service. For example, the 2001 female population of the Air Force is approximately 18.6% while 6.2% of the Marine Corps population are female. The CSM for the i[th] PSU is as follows:

$$C_{hi} = \sum_{k} f_{hk} M_{hik}$$

where,

f_{hk} is the sampling rate for the k[th] second-stage stratum in the h[th] first-stage stratum; and

M_{hik} is the frame counts of persons within PSU *i* for the k[th] second-stage stratum and the h[th] first-stage stratum.

Table 1 provides an example of CSMs calculated for installations (PSUs) from the *1998 DoD* sampling frame. This example demonstrates the usefulness of using a CSM to oversample women by giving PSU *B* a higher probability of selection over PSU *C* due to the percent of females at this installation instead of the overall count of personnel.

Table 1. Example of Composite Size Measures for Four Installations.

PSU	CSM	Pop	Females
A	253.3	13,014	14.7%
B	212.5	7,862	34.3%
C	205.4	10,660	10.4%
D	178.8	8,088	22.3%

4. OPTIMIZATION

We chose a non-linear optimization procedure to determine the number of PSUs and SSUs to select for the *1998 DoD Survey* to satisfy the Kuhn-Tucker necessary conditions. If we denote the precision requirement for the sample proportion from the d^{th} domain as V_d, the sample allocation problem is then formulated as minimizing the cost function *C* from above subject to the following constraints:

$$Var\left(\bar{y}_d\right) \le V_d,$$
$$0 \le x_{\min,n_h} \le n_h \le x_{\max,n_h}, \text{ and}$$
$$0 \le x_{\min,m_{hk}} \le m_{hk} \le x_{\max,m_{hk}}.$$

where,

$x_{\min,n_h}, x_{\max,n_h}$ are the lower and upper bounds on the number of PSUs to be selected per first-stage stratum *h*; and

$x_{\min,m_{hk}}, x_{\max,m_{hk}}$ are the lower and upper bounds on the number of SSUs to be selected per first-stage stratum *h* and second-stage stratum *k*.

The variance components and the variance constraints were estimated from data collected in the *1995 DoD Survey*. To provide stable estimates, three groups of prevalence estimates were used in the optimization routine.

Table 2. Outcome Groups Used in the Calculation of Variance Constraints

Outcome Group	Outcome Category
Drug Use	Marijuana Use
	Any Drug Except Marijuana
	Any Drug Use
Tobacco Use	Any Smoking in Past 30 Days
	Heavy Smoking in Past 30 Days
	Smokeless Tobacco Use (Males Only)
Alcohol Use	Attempted to Quit Smoking
	Abstainers
	Infrequent/Light Drinkers
	Moderate Drinkers
	Moderate/Heavy Drinkers
	Any Drinking Versus Abstainers
	Serious Consequences Due to Alcohol
	Productivity Loss Due to Alcohol
	Alcohol Dependence Symptoms

The components used in the variance constraints were calculated by averaging the estimated variance components of the outcome categories within each outcome group. Negative estimates were converted to zero. In addition to the variance constraints, we imposed practical limitations on the allocation (*Table 6)*. For example, we set an upper limit on the number of SSUs (active-duty members) to be selected from

an installation to ensure that the data collection effort would not become unmanageable either for ourselves or the installation commanders.

5. EVALUATION

Data from the *1998 DoD Survey* have been collected and evaluated since our initial paper. *Table 3* provides a comparison of the respondents to the *1995* and *1998 DoD Surveys*. The optimization for the *1998 DoD Survey* specified a minimum of 4,000 female respondents. The survey nearly achieved this specification with 3,968 female respondents. Note that this is an increase of almost 1,000 completed interviews compared with the *1995 DoD Survey*. This increase accounts for most of the overall increase in respondent numbers.

Table 3. Comparison of Respondents to the 1995 and 1998 DoD Surveys

	1995	1998
Army	3,638	5,449
Navy	4,265	3,930
Marines	3,960	3,622
Air Force	4,330	4,263
DoD	16,193	17,264
Males	13,219	13,296
Females	2,974	3,968
DoD	16,193	17,264

In addition, using this optimization method netted a 15% reduction in the *1998* overall median design effect over the previous version of the survey, in spite of an increase in the oversampling of women (Bray et al., 1995; Bray et al., 1998). *Table 7*, located at the end of the paper, provides a comparison of the population and respondent distributions by Service, gender, and rank. The population distribution of women in the active-duty military increased by 1.3% from 1995 to 1998. However, the distribution of female respondents increased by 4.6%. In spite of the increase in oversampling of females and therefore an increase in the overall design effect, the optimization resulted in an increase in efficiency.

The following table (*Table 4*) lists the median design effects from the *1995* and *1998 DoD Surveys* by Service and the percent difference between the two years. The median design effects were calculated using 17 prevalence estimates from questions included in both questionnaires. We believe the slight increase in the median design effect for the Army and the Marine Corps is attributed to the oversampling of females.

Table 4. Comparison of Median Design Effects by Service from the 1995 and 1998 DoD Surveys

	1995	1998	% Diff
DoD	4.2	3.5	-15.2
Army	2.8	2.9	4.3
Navy	3.8	2.5	-35.1
Marines	3.2	3.4	5.5
Air Force	2.2	1.9	-16.5

For brevity, we have included only the percent difference of the median design effects by gender and Service in *Table 5*. Both males and females overall showed more than a 25% decrease in the design effect level. The percent difference varies across Service within gender.

Table 5. Percent Difference of Median Design Effects by Gender and Service from the 1995 and 1998 DoD Surveys

	Males	Females	DoD
DoD	-27.5	-29.3	-15.2
Army	-4.2	-23.6	4.3
Navy	-38.7	-34.9	-35.1
Marines	-1.2	-8.2	5.5
Air Force	-9.2	0.7	-16.5

We examined the percent differences for other demographic characteristics. The variation was most extreme with military rank. We attributed this variation to the difference in response patterns.

Lower-grade enlisted personnel have historically shown lower response rates to surveys in contrast to junior-grade officers.

6. CONCLUSIONS

Our optimization method developed for the *1998 DoD Survey* used a variance decomposition that accounted for clustering at the first stage and stratification at the second-stage. By accounting for this stratification, our formulation provided accurate estimates of the study design effects. Our optimization increased the number of females in the resulting analysis data set allowing for increased power in the comparison of males and females. Additionally, our optimization was effective in reducing the overall median design effect for the *1998 DoD Survey* by 15% and reducing the median design effects for males and females by over 25%. We plan to use this optimization method for future rounds of the DoD Survey and increase the efficiency of the study.

7. REFERENCES

Bray, R.M., L.A. Kroutil, S.C. Wheeless, M.E. Marsden, S.L. Bailey, J.A. Fairbank, and T.C. Harford (1995). "1995 Department of Defense Survey of Health Related Behavior Among Military Personnel", RTI/6019/06-FR.

Bray, R.M., R.P. Sanchez, M.L. Ornstein, D. Lentine, A.A. Vincus, T.U. Baird, J.A. Walker, S.C. Wheeless, L.L. Guess, L.A. Kroutil, V.G. Iannacchione (1998). "1998 Department of Defense Survey of Health Related Behavior Among Military Personnel", RTI/7034/006-FR.

Chong, E.K.P., and S.H. Zak (1996). *An Introduction to Optimization*. John Wiley & Sons, New York.

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Folsom, R.E., F.J. Potter, and S.R. Williams (1987). "Notes on a Composite Size Measure for Self-Weighting Samples in Multiple Domains." *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 792-796.

Hansen, W.H., W.W. Hurwitz, and W.G. Madow (1953). *Sampling Survey Methods and Theory*. John Wiley and Son, New York.

Horvitz, D.G. and D.J. Thompson (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association* 47, 663-685.

Liu, J., V.G. Iannacchione, and J.D. Kavee (1998). "Estimating Variance Components for a Two-Stage Design with Second-Stage Strata Nested Within PSUs." *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 657-661.

Table 6. Design Constraints Used in the Sample Allocation for the 1998 DoD Survey

			Target	Achieved
Constraints on the Number of PSUs				
Min # of PSUs per Stratum >=			2	2.0
Total # of PSUs <=			65	58.5
Max # of PSUs per Service <=			18	15.8
Max # of PSUs for Army OCONUS <=			6	6.0
Max # of PSUs for Navy OCONUS <=			6	6.0
Max # of PSUs for Marines OCONUS <=			2	2.0
Max # of PSUs for Air Force OCONUS <=			4	4.0
Min # of PSUs per Service >=			12	13.5
Constraints on the Number of SSUs				
Max Total SSUs <=			18,000	18,000.0
Min SSUs per Cell >=	Male		2	12.5
	Female		1	1.7
Max SSUs per Cell <=	Male		1,300	1,017.8
	Female		300	300.0
Min # of DoD SSUs	Female		4,000	4,000.0
Min # of SSUs per PSU >=			250	275.0
Max # of SSUs per PSU <=	Army	CONUS	300	300.0
		OCONUS	350	350.0
	Navy	CONUS	300	275.0
		OCONUS	350	350.0
	Marines	CONUS	300	281.1
		OCONUS	350	350.0
	Air Force	CONUS	300	300.0
		OCONUS	350	350.0

Table 7. Comparison of Population and Respondent Distributions from the 1995 and 1998 DoD Surveys

		1995		1998	
		Population	Respondents	Population	Respondents
DoD		1,325,394	16,193	1,352,614	17,264
Service	Army	31.9%	22.5%	33.7%	31.6%
	Navy	28.8%	26.3%	27.7%	22.8%
	Marines	11.0%	24.5%	11.8%	21.0%
	Air Force	28.4%	26.7%	26.9%	24.7%
Gender	Male	87.6%	81.6%	86.3%	77.0%
	Female	12.4%	18.4%	13.7%	23.0%
Rank	Enlisted	84.4%	77.4%	83.7%	75.3%
	Officer	15.6%	22.6%	16.3%	24.7%