OPTIMIZATION IN THE DESIGN OF A LARGE-SCALE STATE SAMPLE

Sidney H. Schwartz, Charles D. Cowan and Kenneth R. Sausman
U.S. Bureau of the Census

I. INTRODUCTION

One of the most important questions concerning the recently conducted 1980 census is "how well did it work, how many people were missed?". A large-scale state and substate sample, entitled the Post Enumeration Survey (PES), was planned to answer these questions.

This survey would, in part, produce estimates of undercount (the number of persons missed) for each of the fifty states and the District of Columbia with a specified coefficient of variation, as well as estimates of undercount for thirty-two selected cities, and for the entire metropolitan area in which these cities lay, called Standard Metropolitan Statistical Areas (SMSA's). The thirty-two selected cities were chosen based on two criteria. First, a size criteria was chosen; all those cities whose population exceeded five hundred thousand in either 1970 or 1976 were included. Second, a minority population criteria was used; all those cities whose 1970 census population exceeded two hundred fifty thousand and whose 1970 minority population exceeded forty percent of the total population for that city were included.

The PES sample design consisted of three main stages; optimization, stratification, and sampling. A general description of these three main stages will now be given. The remainder of this paper will give a rather detailed description of the optimization portion of the survey.

In brief, the optimization dealt with determining for each sampling area of interest (introduced above, and described further below), the number of enumerators to be used, the number of blocks each enumerator would canvass, and the number of housing units to be interviewed per sampled block. The stratification procedure dealt with the stratification of counties in each state into an appropriate number of strata (determined as a by-product result of the optimization), based on several demographic variables which were found, via regression analysis, to correlate fairly well with the undercoverage rate. Some counties would be determined as self-representing; that is, they would comprise their own stratum. The criterion for self-representing counties was based on population size. Finally, the sampling procedure dealt with the selection of two counties from each non-self-representing stratum using Durbin selection, the actual selection of blocks, and, based on an estimate of the block size, a "take-every" figure to be used for the systematic selection of housing units within the sampled blocks. In this context, the connotation of a block is usually that of a well-defined rectangular piece of land bounded by streets or roads. However, it may be irregular in shape or bounded by railroad tracks, streams, or other features.

II. DESCRIPTION OF OPTIMIZATION

As mentioned above, the PES would produce estimates of undercoverage rates for the fifty states and the District of Columbia, thirty-two central cities, and their entire SMSA's. The optimization thus was required for all the constituent parts of these land areas, namely a) the thirty-two central cities, b) the entire balance of the associated SMSA if the SMSA did not cross a state boundary, c) the within-state portions of SMSA's which did cross a state boundary, d) balances of states which contained any of the geographic areas in a), b), or c), and e) entire states, for those states which did not contain any of the land areas in a), b), or c). The one hundred twenty-one geographic areas determined in a) through e) were called Design Geographic Areas (DGA's).

There were many goals intended for the PES. They were:

a) To estimate the total corrected population for each of the fifty states and the District of Columbia with a specified coefficient of variation. Here, "total corrected population" is defined to be the published census count divided by (1 - estimate of undercoverage rate);

b) To obtain equally reliable estimates of total corrected population at the regional and divisional level;

c) To obtain acceptable estimates of corrected Spanish and corrected Black populations at the regional level;

d) To obtain limited estimates, at the national level, of total corrected population for American Indians, and Asian and Pacific Islanders;

e) To obtain separate estimates of total corrected population for the thirty-two selected cities; and

f) To estimate total corrected population for the entire SMSA's of the thirty-two central cities in "e," above.

As a result of the varying levels of precision indicated by the above categories, and the need to limit the total sample size of the survey to 250,000 households, the value of the specified coefficient of variation varied by DGA. However, each (entire) state was expected to achieve an estimated coefficient of variation on total corrected population of .34 percent.

The optimization portion of the PES would not only determine the values of the three variables indicated earlier, but would also choose for most DGA's between two possible sampling schemes, which were called a "one-stage" design and a "two-stage" design. The choice between the two was primarily based upon which gave the minimum cost for a fixed variance on the estimate of undercoverage.

The fixed variance for a particular DGA "A," say, was computed as follows. If c.v. (•) denotes the coefficient of variation of an estimate, and the coefficient of variation on total corrected population for DGA "A" was t, then

$$\text{c.v.}\left(\frac{N_A}{1-P_A}\right) = t \qquad (1)$$

where n_A is the 1970 (uncorrected) population of DGA "A," and P_A is an estimate of undercoverage rate for DGA "A." (Such estimates of undercoverage rates were derived from 1970 demographic estimates.) Equation (1) implies

$$\frac{V\left(\frac{N_A}{1-P_A}\right)}{\left(\frac{N_A}{1-P_A}\right)^2} = t^2 \qquad (2)$$

where V (•) represents the variance of the estimate.

It can be shown (by using Taylor series expansion, for example) that

$$V\left(\frac{N_A}{1-P_A}\right) \simeq N_A^2 \frac{V(P_A)}{(1-P_A)^4} \qquad (3)$$

so equation (2) implies $V(P_A) \simeq t^2 (1-P_A)^2$. Thus the value of the fixed variance for each DGA can be computed upon knowing the c.v. value for the DGA and the estimate of undercoverage rate for the DGA. A description of the two models along with the cost equations used for each model along with the cost equations used for each model are now given.

A. ONE-STAGE DESIGN

The one-stage design in the optimiation involved determining

i) the number of enumerators

ii) the average number of blocks/enumerator

iii) the average number of housing units to be interviewed/block

Contributing factors in our cost model were "start up" costs (training and administrative), travel costs, interviewing costs, and costs for enumerators to list each unit. The original one-stage cost model was:

$$C=n_I c_I+\bar{d}c_M n_I\bar{n}_B(a_1\bar{\bar{k}}+a_2\bar{\bar{k}}^{\alpha}+a_3)+n_I\bar{n}_B\bar{\bar{k}}c_H+n_I\bar{n}_B c_L \qquad (4)$$

Variables appearing in (4) for which the optimization procedure produced optimum values are:

C = total cost

n_I = number of enumerators

$\bar{n}_B$ = average number of blocks per enumerator

$\bar{\bar{k}}$ = average number of housing units to be interviewed per sampled block.

An additional variable appearing in (4) is:

$\bar{d}$ = average miles traveled for a one-way trip from an enumerator's home to a block of assignment. This depends on the number of enumerators assigned to the DGA (n_I), and was originally set equal to

$\frac{1}{2}\sqrt{\frac{A_R}{n_I}}$, where A_R is the area of the DGA in square miles.

Constants appearing in (4) are:

c_I = "Start up" cost per enumerator, including training and administration

c_M = Cost per mile for travel, including cost of driver.

c_L = Cost of listing all the housing units in a block.

c_H = Cost of enumerating one household

a_1, a_2, a_3, α are constraints which allow $a_1\bar{\bar{k}}+a_2\bar{\bar{k}}^{\alpha}+a_3$ to be a suitable non-linear model for the average number of one-way trips made by an enumerator from home to a block of assignment to complete all the interviews in that block.

REMARK: Note that the first term of the cost equation represents the total "start up" cost, the second term, the total travel cost, the third term, the total interviewing cost, and the fourth term, the total listing cost. Also note that the quantity $\bar{n}_B\bar{\bar{k}}$ represents the total average workload for one enumerator, that is, average total number of housing units to be interviewed by one enumerator, and that $n_I\bar{n}_B\bar{\bar{k}}$ represents the total sample size, in households, for the DGA.

The original variance model used for the one-stage design was

$$V = \frac{g}{n_I n_B \bar{\bar{k}}} + \frac{e}{n_I\bar{n}_B} \qquad (5)$$

Here, variables $n_I\bar{n}_B$ and $\bar{\bar{k}}$ are as given earlier, $V=V(P_A)$ is the fixed variance on the estimate of undercoverage, the value of which is found by using equation (3), and constants g and e are given by

$$g = \tfrac{1}{2}P_A(1-\delta_W)$$

$$e = \begin{cases} \tfrac{1}{2}P_A\delta_W + (.0075)^2 & \text{if } P_A < .025 \\ \tfrac{1}{2}P_A\delta_W + \left((.0075)\frac{P_A}{.025}\right)^2 & \text{if } P_A \geq .025 \end{cases}$$

Here, P_A is the estimate of the undercoverage rate for the DGA "A" under consideration, and δ_W is a constant representing the intraclass correlation of households within blocks with respect to containing a missed person. In addition, "g" represents the correlated component of variance between households within blocks with respect to containing a missed person. The amount .025 appearing in the definition of the variable e represents the average national undercoverage rate from the 1970 census.

Originally, the method of Lagrangian multipliers was used to find (provisionally) optimal values for n_I, $\bar{n}_B$, and $\bar{\bar{k}}$; that is, those values which would minimize the total cost C in equation (4) for a fixed variance V in equation (5). It was found that

$$\bar{\bar{k}} = \frac{g((\bar{\bar{k}}c_H+c_L)n_I^{\frac{1}{2}}+n_I^{\frac{1}{2}}\bar{d}c_M(a_1\bar{\bar{k}}+a_2\bar{\bar{k}}^{\alpha}+a_3))}{(g+e\bar{\bar{k}})(n_I^{\frac{1}{2}}\bar{d}c_M(a_1+\alpha a_2\bar{\bar{k}}^{\alpha-1}+c_H n_I^{\frac{1}{2}})} \qquad (6)$$

and

$$\bar{n}_B = \frac{2c_I n_I^{1/2}}{d' c_M(a_1\bar{k}+a_2\bar{k}^{\alpha}+a_3)}, \quad (7)$$

where $d' = \frac{1}{2}\sqrt{A_R}$

is a set of equations which, through iteration, produces optimal values for $\bar{k}$ and $\bar{n}_B$, respectively. However, since equation (7) depends on the unknown value of n_I, optimal values were found in the following manner:

Equation (5) implies $n_I\bar{n}_B = \frac{1}{V}(e+\frac{g}{\bar{k}})$ (8)

Substituting (7) into (8) and solving for n_I gives:

$$n_I = \left(\frac{d' c_M(a_1\bar{k}+a_2\bar{k}^{\alpha}+a_3)}{2c_I V}\left(e+\frac{g}{\bar{k}}\right)\right)^{2/3} \quad (9)$$

Equation (9) was used with equation (6) in an iterative manner to find provisionally optimal values for $\bar{k}$ and n_I. These resulting values were substituted into equation (7) to obtain a value for $\bar{n}_B$.

REMARKS: Notice that the one-stage design model does not require counties being chosen from the DGA. Thus, the stratification procedure and the Durbin selection in the sampling procedure were not necessary for any DGA found to require a one-stage design. The fact that areas covered by enumerators in a one-stage design may cross county borders is reflected in the variance model by the variable "e."

B. TWO-STAGE DESIGN

i) determining the number of enumerators
ii) determining which counties become self-representing
iii) stratifying the non-self-representing counties
iv) selecting a sample of counties from each stratum of non-self-representing counties
v) determining the number of blocks/enumerator for each sampled county
vi) determining an average number of housing units to be interviewed/sampled block

The optimization procedure was concerned with items i, v, and vi in the above outline.

As in the one-stage design, the contributing factors in the two-stage design cost model were "start-up" costs (training and administrative), travel costs, interviewing costs, and listing costs.

The two-stage cost equation was given by

$$C = n_I c_I + n_I\bar{n}_B \bar{d} c_M(a_1\bar{k}+a_2\bar{k}^{\alpha}+a_3) + n_I\bar{n}_B\bar{k}c_H + n_I\bar{n}_B c_L \quad (10)$$

Variables C, n_I, $\bar{n}_B$ and $\bar{k}$, and constants c_I, c_M, c_L, c_H, a_1, a_2, a_3 and α have the same meanings as in the one-stage design.

Variable $\bar{d}$, although still representing the average mileage for a one-way trip from the enumerator's home to a block of assignment, now was defined equal to $\frac{1}{2}\sqrt{A_C}$, where A_C represents the average area of a county in the DGA under consideration. The two-stage definition of $\bar{d}$ was independent of n_I, the number of enumerators, as the two-stage model originally assumed one enumerator per PSU. Note that, as before, the first term of the cost equation represents the total "start-up" cost; the second term, the total travel cost; the third term, the total interviewing cost, and the fourth term, the total listing cost.

The variance model for the two-stage design was slightly more elaborate than that of the one-stage design (as could be expected due to the greater complexity of the model), and was given by:

$$V = \frac{f}{n_I} + \frac{g}{n_I\bar{n}_B\bar{k}} + \frac{e}{n_I\bar{n}_B} \quad (11)$$

As before, $V=V(P_A)$ is the fixed variance on the estimate of undercoverage, variables n_I, $\bar{n}_B$, and $\bar{k}$ are as given before, and constants e, f, and g in the model are defined by:

$$f = \begin{cases} (.0075)^2 & \text{if } P_A < .025 \\ \left((.0075)\frac{P_A}{.025}\right)^2 & \text{if } P_A \geq .025 \end{cases}$$

$$g = \tfrac{1}{2}P_A(1-\delta_w)$$

$$e = \tfrac{1}{2}P_A\delta_w$$

Here P_A and δ_w have the same interpretation as in the variance model of the one-stage design. In this variance model, "f" represents the between PSU variance. "g" and "e" together represent the within PSU variance, "g" indicating the simple component of variance between households within blocks, and "e" representing the correlated component of variance between households within blocks with respect to containing a missed person. As in the one-stage design, the amount .025 appearing in the definition of the variable f represents the average national undercoverage rate from the 1970 census.

The method of Lagrangian multipliers was used to find provisionally optimal values for n_I, $\bar{n}_B$, and $\bar{k}$; that is, those values which would minimize the total cost C in equation (10) for a fixed variance V in equation (11).

It was found that:

$$\bar{k}_{i+1} = \frac{g}{g+e\bar{k}_i}\frac{\left[\bar{d}c_M(a_1\bar{k}_i+a_2\bar{k}_i^{\alpha}+a_3)+c_L+\bar{k}_i c_H\right]}{\left[\bar{d}c_M(a_1+\alpha a_2\bar{k}_i^{\alpha-1})+c_H\right]} \quad (12)$$

is a recursive formula which produces an optimal value for $\bar{k}$. Originally, a stopping criterion of "stop for the smallest j such that $.99\bar{k}_j \leq \bar{k}_{j+1} \leq 1.01\bar{k}_j$" was used.

Using the resulting value of $\bar{k}$,

$$\bar{n}_B = \left\{\frac{c_I}{f\bar{k}}\left[g+e\bar{k}\right]\left[\bar{d}c_M(a_1\bar{k}+a_2\bar{k}^{\alpha}+a_3)+c_L+\bar{k}c_H\right]^{-1}\right\}^{1/2} \quad (13)$$

is found. Finally, equation (11) implies

$$n_I = \frac{1}{V}\left[\frac{e}{\bar{n}_B} + f + \frac{g}{\bar{n}_B\bar{k}}\right] \quad (14)$$

and, by using equations (12) and (13), n_I is thus determined.

REMARKS: Note that, in contrast to the one-stage design, areas canvassed by enumerators in a two-stage design model do not cross county boundaries. This is reflected in the more complex nature of the two-stage model.

C. ASSUMED VALUES OF CONSTANTS

After several discussions concerning these models with members of Statistical Methods, Field, and Population Divisions at the Bureau of the Census, the following preliminary estimates were made concerning the value of the constants appearing in the cost and variance equations of both designs.

Constant:	c_I	c_M	c_L	c_H
Definition:	Start-up cost per enumerator including training and administration	Cost per mile for travel, including cost for driver	Cost of listing all the housing units in a block	Cost of enumerating one household
Original Value:	\$400	\$.50	\$130	\$20

Constant:	a_1	a_2	a_3	α	δ_w
Definition:	constants allowing $a_1\bar{k}+a_2\bar{k}^{\alpha}+a_3$ to be an adequate model for number of one-way trips made by an enumerator				intraclass correlation of households within blocks with respect to con- containing a missed person
Original Value:	1/6	2	0	.025	.1

III. THEORETICAL RESULTS, DIFFICULTIES, AND RECONCILIATION

The theoretically optimal results posed several difficulties with respect to their practical application. Perhaps the greatest difficulty was the greatly varying values of the product $\bar{n}_B\bar{k}$, the average workload for one enumerator. Frequently, the value of this product was greatly in excess of the amount felt to be reasonable for one enumerator.

Other difficulties of a practical nature of the originally optimal results were:

1. Sometimes, very "large" values of n_I appeared for some very "small" DGA's.
2. Minor perturbations of the parameter values did not produce sufficiently acceptable values of the product $\bar{n}_B\bar{k}$.
3. Widely varying values of $\bar{k}$ from DGA to DGA forced the sampling rates of housing units to vary greatly, an administratively undesirable result.
4. Occasionally, one design would produce a much smaller sample size ($n_I\bar{n}_B\bar{k}$) than the alternative design, but would have a higher minimum cost than the alternative. This somewhat unexpected result was counter-intuitive to our expectations.

As a result of these difficulties, the nature of the optimization was reexamined, and the assumptions made of constants were reevaluated. Two major restrictions were added to the optimization in order to produce more applicable results. First, in order to restrict the widely varying values of enumerator workload, the variable $\bar{n}_B$ was fixed at 6. This restriction required a new set of equations to be solved to determine optimal values of n_I and $\bar{k}$. As before, the method of Langrangian multipliers was used, and the following is the resulting equation for the one-stage design:

$$\bar{k} = \frac{g\left(c_L + \frac{c_I}{\bar{n}_B} + \frac{1}{2}\bar{d}c_M(a_2\bar{k}^{\alpha}+a_3)\right)}{a_1\bar{d}c_M(e\bar{k}+\frac{1}{2}g)+\alpha a_2\bar{k}^{\alpha-1}\bar{d}c_M(e\bar{k}+g)} \qquad (15).$$

Equation (5) implies

$$n_I = \frac{e\bar{k}+g}{V\bar{n}_B\bar{k}} \qquad (16)$$

Equations (15) and (16) were used in a dual iteration scheme to produce optimal values, as follows:

1. Starting with a trial value of $\bar{k}$ (say $\bar{k}$=10), compute a value for n_I, using (16).
2. Using this value for n_I, and the present value of $\bar{k}$, compute a new value for $\bar{k}$, using (15).
3. Substitute this value of $\bar{k}$ into (16) to obtain a new value for n_I.
4. Repeat steps 2 and 3, until convergence of the values of n_I and $\bar{k}$ is obtained.

Two-stage design:

An algorithm similar to the one-stage design was applied, and the resulting formula was:

$$\bar{k} = \left(\frac{g\left(\bar{d}c_M(a_2\bar{k}^{\alpha}(1-a)+a_3)+\frac{c_I}{\bar{n}_B}+c_L\right)}{(e+\bar{n}_Bf)(\bar{d}c_M(a_1+\alpha a_2\bar{k}^{\alpha-1})+c_H)}\right)^{\frac{1}{2}}, \qquad (17)$$

After this was iterated sufficiently, the resulting (optimal) value and the fixed value of $\bar{n}_B$ were substituted into equation (11) to obtain

$$n_I = \frac{\frac{e}{\bar{n}_B} + f + \frac{g}{\bar{n}_B\bar{k}}}{V} \qquad (18)$$

as a formula for the optimal value of n_I.

The second restriction placed on the optimization was that, from inspection of the nature of the two sampling designs, and for administrative convenience, all central cities and balances of SMSA's (part or whole) were assigned to use the one-stage design. Further, upon reconsideration of the original values of constants, additional consultations were made with knowledgeable individuals in Field and Statistical Methods Divisions at the Census Bureau, and the following refinements were made of the value and definition of constants.

TWO-STAGE DESIGN

1. The formula for the average miles traveled in a one-way trip from an enumerator's home to a block of assignment was changed from $\sqrt{A_c}/2$ to $\sqrt{A_{c/2}}/2$. This change reflected the fact that "large" self-representing counties required several enumerators, thus reducing the average overall distance traveled by any one enumerator.

2. The definition of "e," the correlated component of variance between households within blocks with respect to containing a missed person, was changed from $\frac{1}{2}P_A\delta_W$ to $.4P_A\delta_W$. This refinement was made as the coefficient of $P_A\delta_W$ should represent 1 divided by the average number of persons per household. Since this average is closer to 2.5, 1/2.5 = .4 is a more accurate coefficient.

3. The definition of "f," the between PSU (county) variance, was revised from

$$f = \begin{cases} (.0075)^2 & \text{if } P_A < .025 \\ \left((.0075)\frac{P_A}{.025}\right)^2 & \text{if } P_A \geq .025 \end{cases}$$

to

$$f = \begin{cases} (.002)^2 & \text{if } P_A < .025 \\ \left((.002)\frac{P_A}{.025}\right)^2 & \text{if } P_A \geq .025 \end{cases}$$

Upon consultations with several individuals, the original formula for between PSU variance was believed to produce an excessively high value relative to the within PSU variance.

ONE-STAGE DESIGN

1. The definition of "e," the correlated component of variance between households within blocks with respect to containing a missed person, was changed from

$$e = \begin{cases} \frac{1}{2}P_A\delta_W+(.0075)^2 & \text{if } P_A < .025 \\ \frac{1}{2}P_A\delta_W+\left((.0075)\frac{P_A}{.025}\right)^2 & \text{if } P_A \geq .025 \end{cases}$$

to

$$e = \begin{cases} .4P_A\delta_W+(.002)^2 & \text{if } P_A < .025 \\ .4P_A\delta_W+\left((.002)\frac{P_A}{.025}\right)^2 & \text{if } P_A \geq .025 \end{cases}$$

This revision was made because "e" in the one-stage design equals the sum of "e" and "f" in the two-stage design.

In subsequent consultations between Statistical Methods and Field Divisions at the Census Bureau, the following values were agreed to be superior to the originally assumed values in reflecting actual costs for both designs.

Constant:	c_I	a_1	a_2	a_3	α
Refined Value:	\$800	1/8	4	4	.2

The value of δ_W remained at .1. The value of c_L, the cost of listing a block, varied by region from a low value of \$375 to a high value (excepting Alaska) of \$475. Alaska, being sparsely populated, was treated separately and a value of \$750 for c_L was assumed for it. The value of c_M, the cost per mile including driver cost, also varied by region from a low value of \$.34 to a high (excepting Alaska) of \$.44 (Alaska, \$1.00). The value of c_H, the cost of interviewing one household, varied from a low of \$20 to a high value of \$23.

The output resulting from these changes proved to be satisfactory for practical use. The value of $\bar{k}$ fell in the range (14.5, 15.5) for all DGA's. This being the case, it was decided to set $\bar{k}$ at 15 in all cases. This decision had two benefits. First, it set an enumerator workload at $\bar{n}_B\bar{k}$ = (6 x 15) = 90 households for each DGA, an amount felt to be reasonable for the time alloted for interviewing.

Second, setting $\bar{\bar{k}}$ equal to 15 was an administrative convenience in determining "take-every" figures for the systematic selection of housing units.

Generally, the "optimal" value of n_I, (number of interviewers), resulting from the optimization was a non-integer value. To determine the best integer value of n_I, the quantities $[n_I]\,\bar{n}_B\bar{k} = [n_I](90)$ and $[n_I+1]\bar{n}_B\bar{k} = [n_I+1](90)$, where $[\cdot]$ is the greatest integer function, were compared to the product $n_I\bar{n}_B\bar{k}$, where n_I and $\bar{k}$ are the actual (non-integer) values from the optimization. On a DGA by DGA basis, a choice was mader whether $[n_I]$ or $[n_I+1]$ should be used.

IV. EPILOGUE

For the two-stage design, after the integer values of n_I, $\bar{n}_B$, and $\bar{k}$ had been determined, the stratification portion of the PES was performed. Self-representing counties were determined, strata were formed from the non-self-representing counties, and two counties were selected from each of these strata using Durbin selection.

The sampling portion of the survey was well under way when unfortunate news was received. The effect of federal budget cuts and some timing problems necessitated the cancellation of the survey in the form presented here. The survey is now being conducted, on a smaller scale, as a supplement to the Current Population Survey, an ongoing survey conducted on a monthly basis by the U.S. Bureau of the Census.

Even though the original PES was cancelled, the theoretical exercise of the optimization was valuable as it showed the inherent potential for conflicts between theoretically optimal results of a survey design and practical application of these results. It also showed that the resolution of these practical difficulties within the assumed model can be quite difficult. Indeed, only after several consultations and reexaminations of the inherent nature of the models were the final estimates of parameter values made.