COMPARISONS OF MULTI-FRAME WITH SINGLE-FRAME SAMPLE DESIGNS USING REGISTRATION AND VOTING SURVEY DATA [1]

Elizabeth T. Huang, Carma R. Hogue, and Cary T. Isaki
U.S. Bureau of the Census

I. Research Objectives

The purpose of this study is to compare the cost efficiencies of a two-frame sample design and a single-frame design to estimate minority voting and registration rates. Data from the 1976 Registration and Voting Survey (RAV) and an independent sample from lists of registered persons were used to estimate population parameters (cost and per unit variances, covariances and level) in 8 jurisdictions. The first phase of the research which was presented in the Survey Research Methods Section of the 1978 ASA meetings (7) described the methodology and some of the preliminary results with respect to the two-frame design.

The analysis in the first phase referred to as the two-frame design was conducted without considering optimum sample size allocations and optimum weights for the multiple-frame estimator. In this portion of the research, we consider five models, denoted A through E, in order to compare several two-frame and single-frame approaches.

The methodologies used for comparison were dictated by the availability of population parameters that could easily be estimated. Other methodologies were available for comparison but such alternatives would have required the calculation of parameters that were not immediately available.

II. Model Specifications

In order to evaluate the cost efficiency of different survey designs that could have been used in the Registration and Voting Survey, five models denoted A through E were studied in estimating the record-checked minority voting or registration rate in each jurisdiction. The record-checked voting and registration characteristic was obtained by verifying the voting and registration status of each individual in the sample versus the county registration and voting lists. Households had to be interviewed so that responses could be classified by minority.

Models A & B: Models A & B use the same two-frame sample design but differ in their assumptions about the completeness of frame II defined below. One frame denoted frame I is presumed to cover the entire target population of persons eligible to register and consists of four mutually exclusive strata. One stratum (HH) consisted of single unit dwellings selected from tne 1970 Census tapes. The second (GQ) consisted of group quarters (clusters of approximately 3 living quarters) also selected from the tapes. The third stratum (NC) consisted of clusters of approximately four living quarters selected from building permits issued since 1970. The fourth stratum (AREA) consisted of area segments of approximately four households per segment. A stratified simple random sample was drawn from frame I. The other frame, denoted frame II, was the county registration list of persons. From frame II a sample of persons was selected. The manner in which persons were selected from the registration list varied among the counties. In Honolulu County, a two-stage stratified design was used in which election precincts, stratified by minority concentration, were used as the first stage units. A sample of persons was then selected by a simple random sample (SRS) within the selected precincts. In the remaining counties, election precincts were stratified by minority concentration and a SRS of persons was selected over all precincts within each stratum. Clearly, frame II is a subset of the target population.

In models A and B, all persons in selected households from frame I were record-checked. While a sample of persons was selected from frame II the estimator used over frame II utilized the household as the reporting unit after weighting the responses inversely by the number of persons in the household whose current name and address were on the registration list. Sample households containing at least one individual that was found on the registration list with the same name and address as reported on the survey questionnaire were termed linkable. Persons sampled from frame II that were not found at the address provided on the registration list were denoted as sample persons moved (SPM) and the household data were not used in subsequent estimation procedures.

Let $\hat{Y}_1$ denote the estimator of total voting over the linkable households in frame II and $\hat{Y}_2$ and $\hat{Y}_3$ denote the estimators of the total voting over the linkable and nonlinkable households respectively in frame I.

Let $\hat{Z}_1$, $\hat{Z}_2$ and $\hat{Z}_3$ represent estimators comparable to $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$ except that they refer to the number of citizens 18 or over.

The estimator of the voting rate P for models A and B by county and by minority group is as follows:

$$\hat{P} = \frac{\alpha \hat{Y}_1 + (1-\alpha)\hat{Y}_2 + \hat{Y}_3}{\beta \hat{Z}_1 + (1-\beta)\hat{Z}_2 + \hat{Z}_3}$$

where $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ remain to be optimally chosen.

The population parameters relating to frame II include the effect of the SPM's under model A but the effect of the SPM's is removed under model B. That is, under model B, an up-to-date registration list is assumed which contains registrants with their current name and address.

Model C: Under this model a single-frame, frame I, is used from which a stratified random sample of households (hh's) is selected and surveyed so as to obtain voting and registration information as well as minority status. The sample of households is then stratified by minority designation of head of household within the original strata and a simple random subsample of households within each stratum is selected and record-checked.

[1] Excerpted from "Report on the Registration and Voting Survey Two-Frame Approach," internal document, Bureau of the Census (8).

Although alternative estimators could have been investigated, cost and time constraints confined us to the following estimator of P for Model C:

$$\hat{P}_C = \frac{\hat{y} + (\tilde{Y}-\tilde{y})}{\tilde{Z}}$$

where

$\hat{y}$, $\tilde{y}$ are the estimated voting totals of the designated minority from the subsample using the record-checked and the reported data respectively;

$\tilde{Y}$, $\tilde{Z}$ are the estimated totals of the designated minority from the initial sample of the reported data for the voting (Y) and the eligible to register (Z) characteristics respectively.

Model D: Model D assumes the same overall survey design as Model C except that no subsampling is conducted. The estimator of P is $(\hat{Y}_2 + \hat{Y}_3)/(\hat{Z}_2 + \hat{Z}_3)$, which is a special case of Model A where $\alpha=0$, and $\beta=0$. The estimator is written in terms of $\hat{Y}_2$ and $\hat{Y}_3$ but it should be understood that unlike Models A and B, no separation by linkability status is necessary for estimation in Model D.

Model E: Model E differs from Model D only in that no record-checking of the survey responses is conducted. In this case the bias of using reported rather than record-checked data is assumed to not adversely increase the mean square error. However, examination of the data indicated that the square of the estimated bias exceeded the variance constraints and hence Model E was excluded from further analysis.

III. Model Comparisons Procedure

Since a cost efficiency criterion was used in the model comparisons, our objective was to find the optimum cost for each model under the same (specified) coefficients of variation (C.V.). We begin with a discussion of the optimization for a single characteristic (say voting for the Black minority), and later consider the optimization for multiple characteristics (voting and registration for each of the minorities of interest). In Model A, the estimator of P is a ratio of weighted estimators;

$$\hat{P} = (\alpha \hat{Y}_1 + (1-\alpha) \hat{Y}_2 + \hat{Y}_3)/(\beta \hat{Z}_1 + (1-\beta) \hat{Z}_2+\hat{Z}_3),$$

where the weights for the numerator $(\alpha, 0 \leq \alpha \leq 1)$ and the denominator $(\beta, 0 \leq \beta \leq 1)$ are determined so as to minimize the cost subject to a fixed (C.V.) constraint. A typical cost function for Model A is $C = C_{II} + C_I$; where C_{II} is the total sample cost from the registration list frame $(C_{II} = \sum_i C_{Li}\, n_{Li})$ and C_I is the total sample cost from the household frame $(C_I = \sum_i C_{Hi}\, n_{Hi})$. The variance of $\hat{P}$ can be expressed as

$$V(\hat{P}) = \frac{1}{(E\hat{Z})^2} \{\sum_i \frac{N_{Li}^2}{n_{Li}} (1 - \frac{n_{Li}}{N_{Li}})\, S_{Li}^2\, (\alpha,\beta) + \sum_i \frac{N_{Hi}^2}{n_{Hi}} (1 - \frac{n_{Hi}}{N_{Hi}})\, S_{Hi}^2\, (\alpha,\beta)\}$$

where the first term is the variance from the sample selected from frame II and the second term is the variance from tne sample selected from frame I. Note that the stratum variance from either frame is a convex function of α,β. The usual Neyman allocation gives the optimum stratum sample size. According to Neyman allocation, the optimum sample sizes and costs for each given α,β (α = 0.0(0.1)1 , β = 0.0(0.1)1)were computed using the estimated population parameters from the 1976 RAV Survey. The optimum (α^*,β^*) were selected to minimize cost over all (α,β). Since Model D is a special case of Model A where $\alpha=0$, $\beta=0$, Model A possesses the minimum cost between Models A and D if the optimum α and β differ from 0 or 1.

For Model B, the procedure to obtain the optimum weights (α,β) and sample sizes is the same as Model A, except that the estimated parameters (cost and per unit variances and covariances) were adjusted to reflect a registration list with no SPM. Recall that Model C assumes the use of a double sampling plan with a difference estimator. Only a subsample of households is record-checked. A Neyman allocation of the sample sizes was derived for Model C.

Seven specified variance constraints were chosen for the analysis with C.V.'s for $\hat{P}$ specified to be from 0.04 to 0.1 with increments of 0.01. This enabled the comparison of the optimum costs under each model for several variance constraints. In the analysis of Model A the rather large proportions of sample persons moved (SPM) in the registration lists are retained. Hence, the cost per HH of the registration list frame, the variance and population size parameters from the registration list sample differ from those used in the analysis of Model B which simulates a situation in which there were no SPM's in the registration list. In computing the optimum allocations for all of the models, a restriction that the allocated stratum sample size (n_j) not exceed the population stratum size (N_j) was imposed. Hence, if $n_j > N_j$ we set $n_j = N_j$, and repeat the minimization process omitting the previously mentioned stratum j.

The resulting optimum sampling costs for each model and each single characteristic (the specified minority voting or registration) were obtained for all 8 counties. Since the optimum cost was quite different for each of the several characteristics, the overall analysis of the model comparisons cannot be completed without a further choice of criterion. For example, we could either select a primary characteristic (say voting of the smallest minority in the county) or treat the entire problem as a multiple characteristic optimization. Since both minority voting and registration are of major interest, and since in most counties, there are more than one minority of interest, the latter approach was used for the overall analysis.

In the following, we consider an optimum sample allocation scheme based on multiple characteristics. For Models A and B, the weights (α,β) derived from the single characteristic optimization were used. The typical problem for Models A and B under multiple variance constraints is the minimization of the cost C

$$C = \Sigma\, n_{Li}\, C_{Li} + \Sigma\, n_{Hi}\, C_{Hi}$$

with respect to n_{Li}, n_{Hi} such that

$$V(\hat{P}_{j,y}) \leq V^0_{j,y} \qquad j=1, \ldots, J$$

$$V(\hat{P}_{j,x}) \leq V^0_{j,x} \qquad j=1, \ldots, J$$

where for the j-th minority voting (Y)

$$V(\hat{P}_{j,y}) = \frac{1}{Z_j^2} \{ \Sigma \frac{N_{Li}^2}{n_{Li}} (1 - \frac{n_{Li}}{N_{Li}}) S^2_{Li} (\alpha_{j,y}, \beta_{j,y}) + \Sigma \frac{N_{Hi}^2}{n_{Hi}} (1 - \frac{n_{Hi}}{N_{Hi}}) S^2_{Hi} (\alpha_{j,y}, \beta_{j,y})\}$$

and $S^2_{Li}(\alpha_{j,y}, \beta_{j,y})$, $S^2_{Hi}(\alpha_{j,y}, \beta_{j,y})$ are convex functions of $\alpha_{j,y}$, $\beta_{j,y}$. A comparable variance formula is used for the registration rate $\hat{P}_{j,x}$. The above multi-characteristic problem for sample size allocation under a single-frame design is a nonlinear programming problem. Computer algorithms for its solution have been derived by Hartley and Hocking (1963), Chatterjee (1966), Causey (1972,) Al-Khayyal and Hodgson (1978). Since the computer program developed by Causey (1972) was available, his program was used for all counties except Honolulu. Honolulu County was excluded because the sampling design applied on the registration frame precluded the use of the computer algorithm.

The procedure used in obtaining the optimum weights in the single characteristic setting is a natural way to obtain the weights for the multi-characteristic problem. Such a procedure using the computer program would be as follows: Two sets of vector variables $\underset{\sim}{n}$ (sample sizes) and $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ are involved. For each fixed vector value $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ where the elements are bounded by 0 and 1, the optimum $\underset{\sim}{n}$, now denoted $\underset{\sim}{n}(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ can be derived using the computer program and the resulting survey cost can be determined. Repeating the procedure for all possible vectors of $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ where the elements of $\underset{\sim}{\alpha},\underset{\sim}{\beta}$ are bounded by 0 and 1 in increments of 0.1, one can eventually find the optimum $\underset{\sim}{n}$, $\underset{\sim}{\alpha},\underset{\sim}{\beta}$. Since the computer program is an iterative procedure, it would require an enormous amount of computer time to obtain the optimum weight vector $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ by the above procedure. A "short cut" procedure was tested on the data from Edgecombe County where we allocated $\underset{\sim}{n}(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ by the computer program for a set of $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ in the neighborhood of $(\alpha^*_{j,y}, \beta^*_{j,y}, \alpha^*_{j,x}, \beta^*_{j,x})$, j=1, ... J , derived previously for each characteristic j. The resulting "best" $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ were close to the $(\underset{\sim}{\alpha}^*,\underset{\sim}{\beta}^*)$ derived previously under the single characteristic optimization problem. The weight vectors $\underset{\sim}{\alpha}=\underset{\sim}{0}$ and $\underset{\sim}{\beta}=\underset{\sim}{0}$ were used for Model D in the computer algorithm.

For Model C (the double sampling plan with a difference estimator) the cost function is a linear function of the initial i-th stratum sample size (n_i) and the substratum sample size (n_{ij}). The typical variance function for any characteristic is a linear function of the reciprocal of the n_i, n_{ij}. Causey's modified sample allocation program provided the optimum sampling cost and sample size for Model C using multi-characteristic constraints.

The optimum survey costs for all of the models were computed using his program for different C.V. constraints (0.04 (0.01) 0.1) for 7 counties. The relative survey cost for all models versus Model C is provided in Table 1 for all 7 counties for C.V. = 0.1. The results indicate that in all of the counties (except Edgecombe), Model C was the best in terms of minimum cost: Model B or D was second. In Edgecombe County, Model B was the best model and Model C was next.

As one would expect, the optimum sampling cost also varied by the different tolerances for all characteristics (all minority voting and registration). In the study, we used C.V. for $\hat{P}_j$ for the j-th characteristic of 0.04 to 0.1 with increments of 0.01. The relative survey costs of different C.V. constraints versus the costs for a C.V. constraint of 0.1 for Model C are presented in Table 2. Over the 7 counties, an increase in cost of 17% to 22% would be required to reduce the C.V. from 0.1 to 0.09 and an increase in cost of 39% to 53% to reduce the C.V. from 0.1 to 0.08. For a 0.07 C.V. constraint, the cost is nearly double the cost of the 0.1 C.V. constraint. For C.V. = 0.04, the cost increase is 5 times the cost for a C.V. = 0.1 in Coconino, Edgecombe and Halifax; about 4 times in Pinal, Monroe and Lee; and 3 times in Collier.

In Honolulu County, since the multi-variance constraint sample allocation program could not be applied immediately, an Ad Hoc method was used which allocated the sample in a manner such that all variance constraints were satisfied. The Ad Hoc method consisted of selecting the maximum sample sizes (obtained via the single characteristic variance constraint allocation approach) over all the characteristics considered.

The preferred sampling scheme and relative cost comparisons between the models for Honolulu County are summarized in Table 3. The relative cost comparisons for each model and for different C.V. constraints are tabulated in Table 4. For Honolulu County, Model B is preferred in the C.V. range from 0.04 to 0.08. For C.V. = 0.09 and 0.1 Model B costs 1% more than Model C.

IV. Conclusion

Based on the optimum survey costs for each model, we conclude that:

(1) Model C (the double sampling plan with a difference estimator) was the preferred model with either Model D or B second in preference for all C.V.'s considered. In two counties, Model B was in fact "best" for most C.V. constraints.
(2) Survey costs under a two-frame approach could be reduced by using an updated list frame (compare Model A with Model B).
(3) Given the data on registration and voting and the particular sample designs in this study, the single-frame approach performed better in

terms of cost than the multiple-frame model. Within the single-frame model, a double sampling approach using a difference estimator performed better than a scheme requiring 100% record-checked data.

Acknowledgements

We with to acknowledge the contributions of Quentin Ludgin who provided computational assistance and Henry Woltman who devised the record checking procedures.

REFERENCES

(1) Al-Khayyal, F. A., Hodgson, T. J. and Capps, G. D. (1978), "A Lagrangian Dual Approach for Solving Structured Geometric Program Arising in Sample Survey Design", presented at the national meeting of the Operations Research Society of America, The Institute of Management Sciences, Los Angeles, California, 13-15.

(2) Causey, B. (1972), "Optimal Allocation in Stratified Sampling with Multiple Variance Constraints", Proceedings of the Business and Economic Statistics, 258-262.

(3) Chatterjee, S. (1966), "A Programming Algorithm and Its Statistical Application", Technical Report No. 1, Harvard University.

(4) Hartley, H. O. (1962), "Multiple Frame Surveys", Proceedings of the Social Statistics Section of American Statistical Association, 203-206.

(5) Hartley, H. O. and Hocking, R. R. (1963), "Convex Programming by Tangential Approximation", Management Science, Vol. 9, 600-612.

(6) Hartley, H. O. (1974), "Multiple Frame Methodology and Selected Applications", Sankhya: C, No. 36, 99-118.

(7) Henning, C. R., Woltman, H. F. and Isaki, C. T. (1978), "An Application of Multi-Frame Methodology and Measurement Error Research for the 1976 Registration and Voting Survey", Proceedings of the Survey Research Methods Section of the American Statistical Association, 542-555.

(8) Huang, E. T., Hogue, C. R., Isaki, C. T. (1980), "Report on the Registration and Voting Survey Two-Frame Approach", Internal document, Bureau of the Census.

(9) Lessler, J. T. (1974), "A Double Sampling Scheme Model for Eliminating Measurement Process Bias and Estimating Measurement Errors in Surveys", Department of Biostatistics, University of North Carolina, Institute of Statiistics Mimeo Series No. 949

Table 1. Relative Survey Costs for Each Model vs. Model C for a C.V. of 0.1

County	Cost of Model C	A:C	D:C	B:C
Coconino	$ 9,305	1.16	1.16	1.15
Pinal	24,479	1.23	1.23	1.18
Collier	61,857	1.13	1.09	1.12
Monroe	15,839	1.21	1.21	1.21
Edgecombe	6,369	1.04	1.04	0.93
Halifax	6.763	1.06	1.06	1.08
Lee	13,479	1.12	1.12	1.12

Table 2. Relative Survey Costs of Different C.V. Constraints versus 0.1 C.V. Constraints for Model C

C.V / County	0.04	0.05	0.06	0.07	0.08	0.09	0.10	Cost of Model C for C.V. = 0.1
Coconino	5.06	3.53	2.57	1.95	1.52	1.22	1.	$9,305
Pinal	4.17	3.11	2.38	1.86	1.48	1.21	1.	24,479
Collier	3.24	2.54	2.04	1.68	1.39	1.17	1.	61,857
Monroe	4.16	3.10	2.37	1.85	1.48	1.21	1.	15,839
Edgecombe	5.26	3.61	2.61	1.97	1.53	1.22	1.	6,369
Halifax	5.01	3.50	2.56	1.94	1.52	1.22	1.	6,763
Lee	3.96	3.00	2.32	1.83	1.47	1.20	1.	13,479

Table 3. The Preferred Sampling Scheme and Relative Cost Comparisons between Models (Ad Hoc Method) in Honolulu County

Model \ C.V.	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Model A vs. C	.97	.98	.99	1.04	1.04	1.07	1.12
D vs. C	1.22	1.22	1.22	1.23	1.22	1.22	1.23
B vs. C	.90	.90	.91	.96	.96	1.01	1.01
Cost for Model C	$126,268	81,960	57,357	42,302	32,479	25,690	20,831
Model A vs. D	.80	.80	.81	.85	.85	.87	.91
B vs. D	.74	.74	.74	.78	.79	.83	.82
Cost for Model D	$154,498	100,286	70,178	51,819	39,746	31,455	25,539
Model B vs. A	.93	.92	.92	.92	.92	.95	.90
Preferred Sampling Scheme	B	B	B	B	B	C	C

Table 4. Relative Cost Comparisons of Different C.V. Constraints versus C.V. of 0.1 for Each Model (Ad Hoc Method) for Honolulu County

Model \ C.V.	0.04	0.05	0.06	0.07	0.08	0.09	0.10	Survey Cost for C.V. = 0.1
Model A	5.29	3.44	2.44	1.90	1.46	1.18	1.	$23,230
D	6.05	3.93	2.75	2.03	1.56	1.23	1.	25,539
B	5.42	3.51	2.48	1.93	1.49	1.24	1.	21,002
C	6.06	3.93	2.75	2.03	1.56	1.23	1.	20,831