VOTING RIGHTS SURVEY

Grant Capps, U.S. Bureau of the Census

As is well known, the design of any survey will generally involve certain assumptions and "guesstimates" regarding various unknown parameters (frequently costs and variances). The accuracy and amounts of these assumptions will usually depend upon the available funds, lead-time, and prior information. As this paper demonstrates for the 1976 Voting Rights Survey, designed and conducted by the U.S. Bureau of the Census, there were sufficient resources available so as to reduce the usual educated guesstimating associated with efficient survey design.

I. INTRODUCTION

A. Survey Background. The 1976 Voting Rights Survey was concerned with measuring the voting participation rates for certain minorities in specified jurisdictions scattered across the nation. Congress, the Department of Justice, and the Census Bureau jointly identified 93 jurisdictions to be surveyed. These jurisdictions consisted of 11 towns, 73 counties, and 9 States. The minorities, which varied by jurisdiction, included the Black, Spanish, American Indian, Japanese, Chinese, Filipino, and Native Alaskan ethnic groups. Depending upon the costs involved, either a complete census or sample survey was conducted within each jurisdiction. Enumeration occurred within 6 months of the November 1976 presidential election. The results of the survey are expected to be available by November 1977.

B. Purpose of Paper. The purpose of this paper is to describe the major aspects and considerations involved in the sample design for the 1976 Voting Rights Survey. In doing so, the necessary theory will be developed along with the assumptions involved, and the relevant results will be given. The following major survey design problems and their solutions will be presented at length:

1. The determination of the increase in the variance of the estimated minority voting rate due to the clustering of people within households.
2. The determination of both (a) the increase in the variance due to the clustering of housing units, and (b) the optimum cluster size.
3. For each statewide jurisdiction, the determination of a variance function explicitly denoting the components of variance due to (a) the selection of primary sampling units (PSU's), usually counties, and (b) the subsampling within the chosen PSU's.
4. For each statewide jurisdiction, the joint determination of the optimum combination of within PSU sample size, number of sample PSU's, PSU measure of size, and (within PSU) cluster size.

Certain other relatively straightforward aspects of the sample design, such as the allocation of the sample to the various strata, will also be discussed, but to a lesser extent.

C. Survey Requirements. The survey was designed so that the estimated minority voting rate within each jurisdiction would have about a 10% coefficient of variation (CV). For each identified jurisdiction, all minorities which comprised 5% or more of the 18+ population in the jurisdiction were, by definition, minorities of interest. In those jurisdictions with more than one minority of interest, the 10% CV reliability requirement was applied separately to each such minority. In 30 jurisdictions (all the towns and 19 counties), the estimated cost for a complete census was less than that of a comparable sample survey designed to meet the 10% CV requirement. Thus, in these 30 census jurisdictions, the estimates will be free of sampling error, although nonsampling error will be present.

II. INTRACLASS CORRELATIONS AND DESIGN EFFECTS FOR WITHIN COUNTY SAMPLING

In order to determine the approximate sample size needed to meet the 10% CV reliability requirement in the designated counties, it was first necessary to estimate the variance effects of clustering for both persons and housing units. This section reviews the relevant theory and describes the method with which it was applied in determining an appropriate variance model for sampling within the county jurisdictions. As will be shown later, the conclusions arrived at in this section will also be employed when developing the variance formulae pertaining to the statewide jurisdictions.

A. Notation and Definitions. Consider the following situation in a typical county jurisdiction. Let there be M clusters (primary units) of housing units (listing or secondary units), with the i^{th} cluster containing N_i housing units (HU's) for a total of $N = \sum_{i}^{M} N_i$ HU's. The j^{th} HU in the i^{th} cluster contains K_{ij} people 18 and over (elementary units) for a total of

$K_i = \sum_{j}^{N_i} K_{ij}$ 18+ persons in the i^{th} cluster and

$K = \sum_{i}^{M} K_i$ 18+ persons in the entire county.

Let $\bar{N} = N/M$ and $\bar{\bar{K}} = K/N$. The sampling plan we wish to consider involves the selection of m sample clusters (primary units) followed by the secondary selection of $n_i (\leq N_i)$ sample HU's in the i^{th} selected cluster. Let k_{ij} denote the number of 18+ sample persons in the j^{th} sample HU of the i^{th} sample cluster. Assume that simple random sampling is used at both stages and further, that the second stage sampling fraction $f_2 = n_i/N_i$ is constant for all i. The expected total sample size is $n = E\left[\sum_{i}^{m} n_i\right] = E\left[\sum_{i}^{m} f_2 N_i\right] = m\bar{N}f_2$ HU's and the average number of sample HU's per sample cluster is $\bar{n} = \frac{n}{m} = \bar{N}f_2$ HU's. All persons within a sample HU will of course be interviewed and thus $k_{ij} = K_{ij}$. The expected sample of 18+ sample people is $k = E\left[\sum_{i}^{m}\sum_{j}^{n_i} k_{ij}\right] = E\left[\sum_{i}^{m}\sum_{j}^{n_i} K_{ij}\right] = n\bar{\bar{K}}$ and the average number of 18+ sample people per sample HU is $\bar{\bar{k}} = k/n = \bar{\bar{K}}$.

Let:

$$Y_{ij\ell} = \begin{cases} 1 \text{ if the } \ell^{th} \text{ person in the } j^{th} \text{ HU of cluster } i \text{ is an 18+ minority of interest citizen.} \\ 0 \text{ if not.} \end{cases}$$

and

$$X_{ij\ell} = \begin{cases} 1 \text{ if } Y_{ij\ell} = 1 \text{ and the } ij\ell^{th} \text{ person voted.} \\ 0 \text{ if not.} \end{cases}$$

Population Totals: $Y_{ij} = \sum_{\ell}^{K_{ij}} Y_{ij\ell}$, $Y_i = \sum_{j}^{N_i} Y_{ij}$, $Y = \sum^{M} Y_i$,

Population Means: $\bar{\bar{Y}}_{ij} = \frac{Y_{ij}}{K_{ij}}$, $\bar{\bar{Y}}_i = \frac{Y_i}{N_i}$, $\bar{Y} = \frac{Y}{M}$, $\bar{\bar{Y}} = \frac{Y}{N}$, $\bar{\bar{Y}} = \frac{Y}{K}$.

Similarly define the corresponding population totals and means for the variate $X_{ij\ell}$. The unknown parameter to be estimated by the sample is the minority voting rate $R = X/Y$.

Similar definitions can be attached to the sample quantities by simply replacing the upper case letters with lower case ones. For example, we have:

$$y_{ij\ell} = \begin{cases} 1 \text{ if the } ij\ell^{th} \text{ sample person is an 18+ minority of interest citizen.} \\ 0 \text{ if not.} \end{cases}$$

$$x_{ij\ell} = \begin{cases} 1 \text{ if } y_{ij\ell} = 1 \text{ and the } ij\ell^{th} \text{ sample person voted.} \\ 0 \text{ if not.} \end{cases}$$

and

$y_{ij} = \sum_{\ell}^{k_{ij}} y_{ij\ell}$, $y_i = \sum_{j}^{n_i} y_{ij}$, and $y = \sum^{m} y_i$.

Unbiased estimators for Y and X are $y' = \frac{N}{n} y$ and $x' = \frac{N}{n} x$, respectively. Thus, to estimate the minority voting rate R, we use the (nearly) unbiased estimator $r = \frac{x'}{y'} = \frac{x}{y}$. And, finally, let us define one more set of variables. Let $U_{ij\ell} = X_{ij\ell} - R Y_{ij\ell}$. Define the population totals and means corresponding to $U_{ij\ell}$ exactly as with $Y_{ij\ell}$ and $X_{ij\ell}$.

That is, $U_{ij} = \sum_{\ell}^{K_{ij}} U_{ij\ell}$, $U_i = \sum_{j}^{N_i} U_{ij}$, $U = \sum^{M} U_i = 0$ and similarly for the various population means (note $\bar{U} = \bar{\bar{U}} = U = 0$).

For the sample quantities we again use lower case letters. Begin with $u_{ij\ell} = x_{ij\ell} - R\, y_{ij\ell}$. Note that unlike $y_{ij\ell}$ and $x_{ij\ell}$, $u_{ij\ell}$ is an unobservable random variable. The sample totals are

$u_{ij} = \sum_{\ell}^{k_{ij}} u_{ij\ell}$, $u_i = \sum_{j}^{n_i} u_{ij}$, and $u = \sum^{m} u_i = x - Ry$, with obvious definitions for the sample means. An unbiased estimator for U=0 (but certainly not a statistic) is, of course $u' = x' - Ry'$. u' is involved in the Taylorized form for r.

B. Expressions for the Relative Variance of r.

For the above sampling scheme one can refer to nearly any sampling text and obtain the following straightforward approximation for the relative variance of r:

$$V_r^2 = \frac{\mathrm{Var}(r)}{R^2} \doteq \frac{\mathrm{Var}(x' - Ry')}{R^2 Y^2} = \frac{\mathrm{Var}(u')}{X^2} \tag{1}$$

$$= \frac{M-m}{M} \frac{B^2}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{W^2}{m\bar{n}} \tag{2}$$

where,

$$B^2 = \frac{\sum^{M} (U_i - \bar{U})^2}{(M-1)\bar{\bar{X}}^2} \tag{3}$$

$$= \frac{\sum^{M} (X_i - RY_i)^2}{(M-1)\bar{\bar{X}}^2} \tag{4}$$

and

$$W^2 = \frac{1}{N\bar{\bar{X}}^2} \sum_{i}^{M} \frac{N_i}{N_i - 1} \sum_{j}^{N_i} (U_{ij} - \bar{\bar{U}}_i)^2 \tag{5}$$

$$= \frac{1}{N\bar{\bar{X}}^2} \sum_{i}^{M} \frac{N_i}{N_i - 1} \left[\sum_{j}^{N_i} (X_{ij} - RY_{ij})^2 - \frac{1}{N_i} (X_i - RY_i)^2 \right] \tag{6}$$

The first term in (2) is the familiar between-cluster component of relative variance and the second term is the within cluster component which obviously vanishes if there is no cluster subsampling, i.e., if $\bar{n} = \bar{N}$ or $f_2 = 1$.

Since it is desired to express V_r^2 in terms of known or easily guessimated parameters, it is necessary to modify (2). The best reference for accomplishing such a modification is chapter 6, volume 1, of the Hansen, Hurwitz & Madow [3] sampling text. It is stated there (p. 264) that (2) is very nearly equal to

$$V_r^2 = \left[\frac{1-f}{m\bar{n}} V_L^2 \right] \frac{\hat{V}_L^2}{V_L^2} [1 + \delta_L (\bar{n} - 1)], \tag{7}$$

where, $f = \frac{m}{M} f_2 = \frac{n}{N}$, and

$$V_L^2 = \frac{\sum_{i}^{M} \sum_{j}^{N_i} (X_{ij} - RY_{ij})^2}{(N-1)\bar{\bar{X}}^2} \tag{8}$$

$$\hat{V}_L^2 = \frac{M-1}{M} B^2 + \frac{\bar{N}-1}{\bar{N}} W^2 \tag{9}$$

$$= V_r^2 (m = \bar{n} = 1) \tag{10}$$

and

$$\delta_L \doteq \frac{B^2 - \frac{W^2}{\bar{N}}}{\hat{V}_L^2} \text{ (for large M).} \tag{11}$$

The subscript L denotes the listing unit, which here is the HU. The first term in brackets in (7) is the relative variance for a simple random sample of $m\bar{n}$ HU's. The unbracketed middle term of (7) is a factor which should be just slightly greater than unity and is present only if N_i varies from cluster to cluster. Finally, δ_L is the intraclass or intraclass correlation among HU's within clusters of HU's. δ_L is a measure of the homogeneity or similarity among HU's in the same cluster and satisfies $-\frac{1}{\bar{N}-1} \leq \delta_L \leq 1$, taking on the value one when there is perfect within primary unit homogeneity.

V_L^2 in the first term in brackets of (7) can be eliminated by simultaneously applying both (2) and (7) in the case of a single stage simple random sample of mn=n HU's and equating the results. Using form (2) for a random sample of n HU's yields:

$$\frac{N-n}{N}\,\frac{\sum_i^M \sum_j^{N_i} (X_{ij}-RY_{ij})^2}{n\,\bar{\bar{X}}^2(N-1)} = \frac{1-f}{n}\,V_L^2, \qquad (12)$$

and applying expression (7) gives

$$\left[\frac{1-\frac{n}{N}\frac{\bar{\bar{k}}}{\bar{\bar{K}}}}{n\bar{\bar{k}}}\,V_2^2\right]\frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{k}}-1)]$$

$$=\left[\frac{1-f}{n\bar{\bar{K}}}\,V_2^2\right]\frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{K}}-1)], \qquad (13)$$

where

$$V_2^2 = \frac{\sum_i^M \sum_j^{N_i} \sum_\ell^{K_{ij}} (U_{ij\ell}-\bar{\bar{U}})^2}{(K-1)\,\bar{\bar{X}}^2} \qquad (14)$$

$$\doteq \frac{1-R}{R\,\bar{\bar{Y}}} \qquad (15)$$

$$B_2^2=V_L^2,\quad W_2^2=\frac{1}{K\bar{\bar{X}}^2}\sum_i^M \frac{N_i}{K_{ij}-1}\sum_j^{K_{ij}}\sum_\ell (U_{ij\ell}-U_{ij})^2, \text{ and}$$

$$\hat{V}_2^2 = \frac{N-1}{N}\,B_2^2 + \frac{\bar{\bar{K}}-1}{\bar{\bar{K}}}\,W_2^2$$

$$\delta_2 \doteq \frac{B_2^2-\frac{W_2^2}{\bar{\bar{K}}}}{\hat{V}_2^2} \text{ (for large N).} \qquad (16)$$

The terms in (13) have an interpretation very similar to the corresponding terms in (7). The first term in brackets in (13) is the relative variance for a simple random sample of $n\bar{\bar{K}}$ people. The unbracketed middle term of (13) is a factor which represents the increase in the variance due to K_{ij} varying from HU to HU. δ_2 is the intraclass or intrahousehold correlation among people within households. Equivalently, δ_2 is a measure of the homogeneity among people in the same HU. Equating (12) and (13) yields the following expression for V_L^2:

$$V_L^2 = \left[\frac{V_2^2}{\bar{\bar{k}}}\right]\frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{k}}-1)] \qquad (17)$$

Substituting (17) in (7) and using (15) gives:

$$V_r^2 = \left[\frac{1-f}{mn\bar{\bar{K}}}\,V_2^2\right]\frac{V_2^2}{V_2^2}\,\frac{\hat{V}_L^2}{V_L^2}\,[1+\delta_L(\bar{n}-1)][1+\delta_2(\bar{\bar{k}}-1)] \qquad (18)$$

$$=\left[\frac{1-f}{mn\bar{\bar{K}}}\,\frac{1-R}{R\pi}\right]\mathrm{Def}_L\,\mathrm{Def}_2 \qquad (19)$$

where,

$$\mathrm{Def}_L = \frac{\hat{V}_L^2}{V_L^2}\,[1+\delta_L(\bar{n}-1)] \qquad (20)$$

and

$$\mathrm{Def}_2 = \frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{k}}-1)] \qquad (21)$$

are the design effects for HU's within primary units and for people within HU's, respectively, and where $\pi = \bar{\bar{Y}}/K$ is the fraction of the population in the subgroup of interest. Assuming both design effects and their components can be approximated, the relative variance of r as given in (19) is finally in a desirable and usable form.

C. Estimating the Intraclass Correlations and Design Effects. We now turn to the estimation of the needed parameters. Data from the Current Population Survey (CPS), designed and conducted by the Census Bureau, were employed. The CPS [5] is a nationwide multi-stage sample survey conducted each month with a total sample size of about 56,000 designated HU's. Each election year, both presidential and nonpresidential, a supplement is added to the November CPS questionnaire which contains citizenship, registration, and voting questions. Although the November 1974 data were available, the November 1972 data were used in approximating the unknown parameters. The 1972 data were chosen for two important reasons. First of all, 1972 was a presidential election year, as was the election for which the survey was being designed.

Secondly, and quite fortunately, the 1972 sample consisted of a mixture of two sample designs. Half of the sample was the result of an older design for which $\bar{N}$=18, $\bar{n}$=6, and f_2=1/3. The other half of the sample stems from the CPS redesign and features $\bar{N}$=$\bar{n}$=4 and f_2=1. Having a reading for δ_L in both designs would indicate how δ_L (which is dependent upon $\bar{N}$) varies with changing $\bar{N}$. Only those counties which were self-representing (single counties or groups of counties are the primary sampling units in the CPS) in both designs were used in the estimation.

The total sample size for the study was approximately 20,000 HU's, about 10,000 sample HU's each for the old and new CPS sample designs. The analysis had several features. First of all, the counties in the analysis were placed, on the basis of geographic proximity, in one of four groups or universes and a fifth or combined universe which consisted of every county. The four groups consisted of counties in the Northeast, North Central, Southern and Western regions. Within each CPS sample design, the sample size was about 2,500 HU's in each of the first four universes. Second, since the only race designations collected in the 1972 CPS were White, Black, and Other, these three races, along with a fourth domain which included All races, were each used separately in the analysis. This resulted in 2x5x4=40 (number sample designs x number universes x number races) separate readings on the intraclass correlations and design effects. Unfortunately, due to small sample sizes for both the Black and Other racial subgroups, the readings obtained for these subgroups were considered highly suspect and consequently were of little use. Thus, the only reliable results were those obtained from the racial subgroups of White and All.

The usual consistent estimators for the unknown population parameters were employed in arriving at the following useful approximations for the parameters:

HU's

$$\bar{N}=4:\begin{cases}\delta_L = .166\\ \dfrac{\hat{V}_L^2}{V_L^2} = 1.05\\ \text{Def}_L = 1.05\ [1+.166(\bar{n}-1)]\end{cases} \quad (23)$$

$$\bar{N}=18:\begin{cases}\delta_L = .144\\ \dfrac{\hat{V}_L^2}{V_L^2} = 1.05\\ \text{Def}_L = 1.05\ [1+.144(\bar{n}-1)]\end{cases} \quad (24)$$

Persons

$$\bar{\bar{K}}=2:\begin{cases}\delta_2 = .627\\ \dfrac{\hat{V}_2^2}{V_2^2} = 1.15\\ \text{Def}_2 = 1.15\ [1+.627(\bar{\bar{K}}-1)]\end{cases} \quad (25)$$

Of course, the two underlying assumptions regarding the above conclusions are that (1) the minority and majority are fairly similar with respect to the above parameters, and (2) the counties in the actual survey are not unlike those in the CPS study. If one subscribes to the familiar model $\delta_L=a\bar{N}^b$ and uses the above results for $\bar{N}=4$ and $\bar{N}=18$, to solve for a and b, the obtained solution is,

$$\delta_L=(.1892)(\bar{N})^{-.0945} \quad (26)$$

which clearly demonstrates the dependence of δ_L upon $\bar{N}$. Likewise, δ_2 depends upon $\bar{\bar{K}}$, but since $\bar{\bar{K}}$ varied only slightly (about 2) among the jurisdictions in the actual Voting Rights Survey, equations (25) were considered valid for all jurisdictions (i.e., all $\bar{\bar{K}}$). Thus, the variance function (19) becomes:

$$V_r^2 = \left[\frac{1-f}{mn\bar{\bar{K}}}\ \frac{1-R}{R\ \pi}\right] 1.05[1+\delta_L(\bar{n}-1)]\ 1.15[1+.627(\bar{\bar{K}}-1)] \quad (27)$$

where δ_L depends upon $\bar{N}$ as discussed above. In the actual design of the survey, the value of π for a given jurisdiction was approximated by the 1970 census value and the value of R (obviously unknown) was taken to be the smaller of the 1972 overall voting participation rate for the entire jurisdiction (as given by Richard Scammon's "American Votes" [4] series) and the regional (in some cases national) minority of interest voting rate as estimated by the 1972 CPS. The values of π varied widely from .05 in some jurisdictions to a maximum of about .60, while the assumed minority voting rate generally satisfied $.20 \leq R \leq .40$.

III. SAMPLING FRAMES, COST CONSIDERATIONS AND OPTIMUM CLUSTER SIZES, ALLOCATION OF THE SAMPLE, AND DETERMINING SAMPLE/CENSUS STATUS FOR THE TOWN AND COUNTY JURISDICTIONS

This section will present several very closely related topics in the town and county jurisdictions. The various sampling frames frequently used by the Census Bureau to select general population samples will be described, as will the associated advantages, restrictions and costs for each frame. These costs, along with the already determined variance function, determine the approximate optimum cluster size in the various frames.

Also discussed will be the allocation of the sample to the various frames and strata. This section will conclude with some brief remarks concerning the determination of the sample versus census status of each jurisdiction.

A. Sampling Frames. There are three basic sampling frames which are used by the Census Bureau to select general population samples. A short description now follows for each of these three frames.

1. Old Construction Frames-1970 Census Detail Files. These are a group of files consisting of a detailed record for each, or a subset thereof, April 1970 housing unit. These files are a result of the 1970 census. The files that contain only a subset of the census units are the result of a sample and contain more detailed information for a given unit then does the complete tape. One large advantage of sampling from any of these files is the ease with which a high degree of stratification, based upon 1970 characteristics, is achieved. Of course, due to the movement of the population, the effectiveness of any stratification based upon 1970 characteristics decreases with time. Since units existing prior to April 1970 are referred to as old construction units, the above set of files will be referred to as old construction sampling frames.

2. New Construction Frame-Building Permits. Many counties require and maintain records of all newly constructed inhabitable structures in part or all of the county. These records generally take the form of building permits and contain the number of new HU's existing within the structure. Thus, with the aid of building permits, new HU's built in the permit issuing portions of a county can be sampled. Only a limited amount of stratification can be achieved, however, when sampling from building permits. Unfortunately, the permit issuing portion of a county may either be very small or nonexistent and thus, this building permit frame is often not available. Defining new construction units as those built since April 1970 clearly makes the building permit frame a new construction one.

3. Old and New Construction Frame-Area Maps. Another type of sampling that is widely used at the Bureau is area segmenting and sampling. The sampling frame used in area sampling is a land map showing the 1970 census count of HU's in small land areas. Each small land area (i.e., cluster) contains about twenty (i.e., $\bar{N}=20$) HU's, however, there is a fair amount of variation among these cluster sizes. These area segments are generally sampled with probability proportional to their size (i.e., 1970 HU count) and then subsampled as desired. The achievable degree of stratification is minimal with this type of area sampling, and further, as time goes on, the measures of size become poorer and poorer due to additions and losses of HU's. The advantage of the area frame is the ability to assign positive probabilities of selection to units built after the 1970 census, thus providing an alternative to sampling from building permits which are often unavailable. In addition, the area frame is also

used to sample old construction whenever census addresses from the old construction frame are poor. The area frame is obviously an old and new construction sampling frame.

B. Cost Considerations and Optimum Cluster Sizes. The determination of an appropriate cost model to be used in approximating optimum cluster sizes is often as important and as difficult as the derivation of the variance function. For this survey the importance of the variance function was probably greater than that of the cost function simply because of the strict reliability requirement. Since the emphasis in this paper is on the variance function and its detailed determination, we will sometimes be content with a fairly macroscopic discussion of some of the cost considerations.

In relative terms, it is generally more expensive to sample from the area frame than from either of the remaining two frames, which are each about equally expensive. Thus, it is desirable, from a cost standpoint, to use the 1970 detail files in conjunction with the building permit frame in the permit issuing portions of a given county. There is little choice but to use the area frame in the nonpermit issuing portions. The determination of the cluster sizes in the three frames will now be discussed.

1. Cluster Sizes in the New Construction Frame. It was decided to use the traditional permit new construction sample design of $\bar{N}=\bar{n}=4$. This type of clustering is frequently used at the Bureau and there was some advantage in being able to use established procedures. Also, a very rough cost analysis indicated this to be reasonably optimum. In addition, the new construction sample was generally a very small fraction of the overall sample, thus reducing the importance of optimality.

2. Cost Model for the Old Construction and Area Frames. The standard three term cost equation was developed and employed within each county jurisdiction. The cost model, derived separately for each of the old construction and area sampling frames within each county, took the following form:

$$c=c_1 m + c_2 \bar{n} m + c_3 \sqrt{mA}\ , \qquad (28)$$

where

c = total variable cost,

c_1 = cost per primary unit or cluster (includes cost of selecting, listing, and subsampling the clusters),

c_2 = cost per secondary unit or HU (includes cost of interviewing, processing, and within primary unit travel),

c_3 = cost per mile of travel between clusters (includes mileage and interviewer wages while traveling), and

A = county land area in square miles.

Without discussing the detailed computation of the actual cost coefficients (i.e., c_1, c_2, and c_3), the following observations are of extreme importance when determining the optimum cluster sizes within each sampling frame:

a) For each of the three frames the third term involving travel costs ($c_3\sqrt{Am}$) is negligible compared to the second term ($c_2\bar{n}m$) due to the small land areas A (often less than 500 square miles) generally encountered and due to the fairly large sample size $\bar{n}m$ (usually at least 500 sample HU's).

b) For the old construction frame the second term ($c_2\bar{n}m$) dominates the first term ($c_1 m$). This claim cannot be made for the area frame.

3. Optimum $\bar{N}$ and $\bar{n}$ in the Old Construction Frame. Assume the entire sample is to come from the old construction frame in a given county, subject to meeting the CV reliability requirement $\sqrt{V_r^2} = .10$, where V_r^2 is given by (27). The objective is to minimize the cost c in (28), while attaining this 10 percent CV. Though there is no control over $\bar{\bar{K}}(=\bar{\bar{k}})$, the cheapest combination of $\bar{N}$ and $\bar{n}(\leq\bar{N})$ can be selected. Although no mathematical solution exists for this particular problem, an iterative solution can easily be found as follows. Using (26) for δ_L as a function of $\bar{N}$, the only unknowns in V_r^2, as displayed in (27), for a given jurisdiction are m, $\bar{n}$, and $\bar{N}$. Specifying a given combination of $\bar{n}$ and $\bar{N}$, subject to $\bar{n}\leq\bar{N}$, one can solve for m using (27) and apply (28) to obtain the cost for this $\bar{n}$, $\bar{N}$, m combination. This procedure was followed for all reasonable combinations of $\bar{n}$ and $\bar{N}$ and the old construction sampling frame cost was recorded each time. As one would expect upon returning to the two comments immediately following (28), the winning combination in each county jurisdiction was $\bar{N}=\bar{n}=1$, or equivalently, a simple random sample of HU's.

4. Optimum $\bar{n}$ in the Area Frame. Assume the entire sample is to come from the area frame in a given county. Unlike the old construction sampling frame, it is not possible to select at will the value of $\bar{N}$ in the area frame. This is due to the nature of the area segmenting, in which the HU cluster sizes are variable and average about $\bar{N}=20$. This frame imposed restriction on $\bar{N}$ is, in some sense, similar to the real world imposed restriction on $\bar{\bar{K}}$, over which we have no control either. The area frame optimization procedure was similar to the old construction one, except only one value of $\bar{N}$ was considered, that value being 20. The cost for all combinations of $\bar{n}$ and m such that $V_r^2=.01$ and $\bar{n}\leq\bar{N}=20$ were determined. The cost efficient cluster size in each county jurisdiction was $\bar{n}=4$ HU's.

5. Optima in the Combined Sampling Frames. The optima just derived pertained to the old construction sampling frame (useful in the 100 percent permit issuing jurisdictions) and to the area frame (useful in the 0 percent permit issuing counties). About ¼ of the county jurisdictions were 100 percent permit issuing and a handful were 0 percent permit issuing. Thus, there were many partially permit issuing counties for which it was necessary to select a sample from each of the three frames. In such counties, it was decided to simply use the already determined optima in the various frames. That is, $\bar{N}=\bar{n}=4$ was used in the new construction frame, $\bar{N}=\bar{n}=1$ in the old construction frame, and $\bar{N}=20$, $\bar{n}=4$ in the area frame. Combining the individual sample frame optima to obtain an overall optima is permissible whenever the between cluster travel costs are relatively negligible (see Cochran [1], p. 289), as they are here.

C. Allocation of the Sample. The next step in the sample design was to efficiently allocate the sample to the various sampling frames or strata. When sampling from the 1970 detail tapes, the old construction frame was divided into two strata, those 1970 HU's with and without a minority of interest head. Thus, altogether there are four strata, the two old construction frame strata, the new construction stratum and the area stratum, to which the sample needed to be allocated. A variance function, similar to the earlier one but applicable to a stratified sample design will now be derived. The notation about to be introduced will be an obvious modification of the earlier notation with the first subscript (h) denoting the stratum rather than the cluster. For example:

Y_h = number of 18+ minority of interest citizens in stratum h, (h=1,2,3,4),

$R_h = \frac{X_h}{Y_h}$ = minority of interest voting rate in stratum h,

y'_h = usual unbiased estimator of Y_h based upon a sample of size n_h from N_h.

$r_s = \frac{\sum^4 x'_h}{\sum^4 y'_h}$ = stratified ratio estimator of R, and

Def_{Lh} = design effect for HU's within clusters in stratum h

$$= \left(\hat{V}^2_{Lh} \Big/ V^2_{Lh}\right)[1+\delta_{Lh}(\bar{n}_h - 1)]$$

$$= \begin{cases} 1.000 \text{ in the two old construction strata (h=1,2)} \\ 1.05\ [1+.166(4-1)] = 1.573 \text{ in the new construction stratum (h=3)} \\ 1.05\ [1+.143(4-1)] = 1.500 \text{ in the area stratum (h=4)} \end{cases}$$

The relative variance, $V^2_{r_s}$, of the stratified ratio estimator r_s is

$$V^2_{r_s} = \frac{Var(r_s)}{R^2} \doteq \frac{1}{X^2} \sum^4 Var(x'_h - R\, y'_h).$$

If the minority voting rate is assumed to be approximately the same in each stratum (probably a reasonable assumption), then $R_h = R$ (h=1,2,3,4) and we have

$$V^2_{r_s} \doteq \frac{1}{X^2} \sum^4 Y_h^2\, Var\left(\frac{x'_h - R_h\, y'_h}{Y_h}\right)$$

$$\doteq \frac{1}{X^2} \sum^4 Y_h^2\, Var(r_h) \doteq \frac{1}{X^2} \sum^4 X_h^2\, V^2_{r_h}.$$

The variance function for $V^2_{r_h}$ has already been derived and is given in (27). Using this result yields:

$$V^2_{r_s} \doteq \frac{1}{X^2} \sum^4 X_h^2 \left[\frac{1-f_h}{n_h \bar{\bar{K}}_h} \frac{1-R_h}{R_h \pi_h}\right] Def_{Lh} Def_{2h}$$

$$\doteq \frac{1-R}{RY^2} \sum^4 (1-f_h) \frac{Y_h N_h}{n_h} Def_{Lh} Def_{2h}$$

And finally, if one assumes $\bar{\bar{K}}_h = \bar{\bar{K}}$ (h=1,2,3,4), then we have:

$$V^2_{r_s} \doteq \left[\frac{1-R}{RY^2} Def_2\right] \sum^4 (1-f_h) \frac{Y_h N_h}{n_h} Def_{Lh}.$$

Since the HU costs in the four strata do not differ by more than a factor of two, a Neyman allocation is approximately optimal. Therefore, the sample was allocated to the four strata so that n_h was proportional to $\sqrt{Y_h N_h\, Def_{Lh}}$. In order to perform this allocation, estimates of Y_h and N_h (these are 1976 parameters and hence unknown) were needed. Based primarily upon the 5-year movement rates between 1965 and 1970 for each county, the known 1970 values of Y_h and N_h, the available estimates of new construction as well as a few other assumptions regarding the expected number of people moving into and out of an area, estimates of Y_h and N_h were made and used in the sample allocation.

D. Sample Vs. Census Jurisdictions. The final topic in this varied section will briefly discuss the determination of the sample and census jurisdictions. Costs and selection methods differ markedly between sample surveys and complete censuses. For example, an interviewed census HU will typically cost about $5.00 while the same HU selected by a sample survey might cost about $25.00. This would imply that whenever the sampling fraction $f = n/N$ is greater than .2, a census would be less expensive. Therefore, after determining the sample size and the corresponding sample survey cost for each town and county jurisdiction, and comparing this to the census cost, the sample and census jurisdictions were easily designated. As previously mentioned, in all 11 towns and in 19 of the 73 counties, it was cheaper to conduct a census.

IV. VARIANCE MODEL AND OPTIMA DETERMINATION IN THE STATE JURISDICTIONS

In 9, mostly southern, States, we were required to select a statewide sample in order to estimate the statewide minority of interest voting rate with a 10 percent CV. These 9 State jurisdictions included Arizona, Alaska, Alabama, Georgia, Louisiana, Mississippi, South Carolina, Texas, and Virginia. Arizona, which had 9 of its 14 counties designated as jurisdictions to be surveyed, was the only State among the 9 containing designated county jurisdictions. This section presents the derivation of the variance and cost functions that were extremely valuable in approximating the optimum combination of within PSU sample size, number of sample PSU's, PSU measure of size, and the within PSU cluster size, for each of the 9 States. Other aspects of the statewide sample designs are also discussed.

A. Variance Function. The first topic is the derivation of the all-important variance function. The goal, as has been the case throughout this paper, was to develop a variance model in terms of known or reasonably estimated parameters. In particular, it was also desired, at some point, to make use of the already determined within county variance model of section II.

The basic sampling plan is the following stratified multi-stage design. With the counties designated as the PSU's, stratify the PSU's, select a sample of PSU's from each stratum with replacement and with probability proportional to some measure of size, and subsample the chosen PSU's by first selecting clusters of HU's and then

subsampling the chosen clusters. The final two stages of selection that follow the first stage selection of PSU's is similar to the earlier within county sampling.

The notation to be used in deriving a variance function for this three-stage design is again an obvious modification of the original notation. Each of the original subscripts is to be shifted two places to the right. The first subscript (h) will now designate the stratum and the second subscript (p) will denote the PSU within the stratum. The third subscript (i) denotes the secondary unit (cluster), the fourth (j) denotes the third stage unit (HU), and the fifth (ℓ) denotes the individual people. For example, this new notation results in the following:

Y_h = number of 18+ minority of interest citizens in stratum h (h=1,2,...,H),

$R_h = \frac{X_h}{Y_h}$ = minority of interest voting rate in stratum h,

Y_{hp} = number of 18+ minority of interest citizens in PSU p of stratum h (p=1,2,...,T_h),

$R_{hp} = \frac{X_{hp}}{Y_{hp}}$ = minority of interest voting rate in PSU p of stratum h,

y'_{hp} = usual unbiased estimator of Y_{hp} based upon a sample of size n_{hp} HU's from the N_{hp} HU's in PSU p of stratum h,

where

H= number of strata in the State, and

T_h = number of PSU's in stratum h.

Also let:

t_h = number of sample PSU's in stratum h,

Z_{hp} = single-draw probabilities or normalized measures of size for PSU p of stratum h such that $\sum_p^{T_h} Z_{hp} = 1$,

y'_h = usual with replacement estimator of Y_h

$$= \sum_p^{T_h} \frac{y'_{hp}}{t_h Z_{hp}},$$

y' = usual stratified estimator of $Y = \sum^H y'_h$,

$r_M = \frac{x'}{y'}$ = multistage ratio estimator of R,

Def_{Lhp} = design effect for HU's within clusters in PSU p of stratum h, and

Def_{2hp} = design effect for people within HU's in PSU p of stratum h.

The relative variance of r_M, $V^2_{r_M}$, is first expressed as:

$$V^2_{r_M} = \frac{Var(r_M)}{R^2} \doteq \frac{Var(x'-R\,y')}{X^2} \doteq \frac{1}{X^2}\sum^H Var(x'_h - R y'_h) \quad (29)$$

Using Durbin's [2] (1953) well-known result concerning the variance of a multi-stage statistic, the general term of (29) can be expressed as:

$$Var(x'_h - R\,y'_h) = Var[E(x'_h - R\,y'_h | PSU's)] + E[Var(x'_h - R\,y'_h | PSU's)]$$

$$= \sum_p^{T_h} \frac{Z_{hp}}{t_h}\left[\frac{X_{hp}-RY_{hp}}{Z_{hp}} - (X_h - RY_h)\right]^2 + \sum_p^{T_h} \frac{1}{t_h Z_{hp}} Var(x'_{hp} - R y'_{hp} | PSU's). \quad (30)$$

The first term in (30) represents the familiar between-PSU variance and the second term the within PSU variance. To simplify the conditional within county variance $Var(x'_{hp} - R\,y'_{hp} | PSU's)$, the earlier results (14) and (18) are applied to the variate $U_{hpij\ell} = X_{hpij\ell} - RY_{hpij\ell}$. Ignoring the finite population correction (fpc) factor, this yields:

$$Var(x'_{hp} - R y'_{hp} | h,p) = K^2_{hp} \frac{\sum\sum\sum (U_{hpij\ell} - \bar{U}_{hp})^2}{(K_{hp}-1)\, n_{hp} \bar{\bar{K}}_{hp}} Def_{Lhp} Def_{2hp}. \quad (31)$$

Notice that in (31) it has been subtly assumed that the design effects for the variates $X_{hpij\ell} - R\,Y_{hpij\ell}$ and $X_{hpij\ell} - R_{hp} Y_{hpij\ell}$ are the same. This seems like a reasonable assumption, as design effects are usually fairly robust and these two variates are quite similar. Upon simplifying (31) we obtain

$$Var(x'_{hp} - R y'_{hp} | h,p) = \frac{N_{hp}}{n_{hp}}\left[X_{hp} + R^2 Y_{hp} - 2R\,X_{hp} - \frac{(X_{hp} - R\,Y_{hp})^2}{K_{hp}}\right] Def_{Lhp}\, Def_{2hp} \quad (32)$$

The variance function can now be assembled and expressed as

$$V^2_{r_M} = \sum_h^H \sum_p^{T_h} \frac{(X_{hp} - R\,Y_{hp})^2}{t_h Z_{hp} X^2} - \sum^H \frac{(X_h - R\,Y_h)^2}{t_h X^2} + \frac{1}{X^2}\sum_h^H \sum_p^{T_h} \frac{N_{hp}}{t_h Z_{hp} n_{hp}}\left[X_{hp} + R^2 Y_{hp} - 2RX_{hp} - \frac{(X_{hp} - RY_{hp})^2}{K_{hp}}\right] x\, Def_{Lhp} Def_{2hp}. \quad (33)$$

The first two terms in (33) is the simplified between-PSU relative variance of (30), with the second term explicitly showing the reduction in the total variance due to the stratification. Believe it or not, if one has an available computer, (33) is in a very usable form. X_{hp} can be estimated using 1972 county voting data from Scammon [4] and adjusting to account for the lower minority voting rates and the change in population between 1972 and 1976. N_{hp}, K_{hp}, and Y_{hp} can be estimated using 1970 census data and adjusting for the change in population between 1970 and 1976. Thus, the only unknowns in (33) are the formation of the strata, t_h, Z_{hp}, $\bar{N}_{hp}$, $\bar{n}_{hp}$, and n_{hp}.

For the moment, to aid in the search for the various optima, the following restrictions are placed upon our sample design:

1. Only one stratum will be formed and t PSU's will be selected with replacement from this single statewide stratum.
2. n_{hp} will be assumed constant for all PSU's and be denoted by W (workload).
3. $\bar{N}_{hp} \equiv \bar{N} = 20$ for all PSU's, primarily because it must equal 20 for any area sample.
4. $n_{hp} \equiv \bar{n}$ will be assumed constant for all PSU's thus $Def_L \equiv Def_{Lhp} = 1.05[1 + .143(\bar{n}-1)]$.
5. $Def_{2hp} \equiv Def_2$ is constant in each PSU.

Restrictions 1 and 2 above will later be lifted. Denoting the only stratum by h=1, the variance function (33) under these restrictions becomes,

$$V^2_{r_M} = \sum_p^{T_1} \frac{(X_{1p}-R\,Y_{1p})^2}{t\,Z_{1p}X^2} + \frac{Def_L Def_2}{X^2\,tW} \sum_p^{T_1} \frac{N_{1p}}{Z_{1p}}\left[X_{1p}+R^2Y_{1p}-2RX_{1p}\right] - \frac{Def_L Def_2}{X^2\,tW} \sum_p^{T_1} \frac{N_{1p}}{Z_{1p}}\left[\frac{(X_{1p}-RY_{1p})^2}{K_{1p}}\right]. \quad (34)$$

The unknowns for which the jointly optimum combination is desired have now been reduced to t, Z_{1p}, $\bar{n}(\leq 20)$, and W. Setting $V^2_{r_M}$ =.01 and specifying the set of basic probabilities or measures of size Z_{1p}, along with any two of t, $\bar{n}$, and W, will determine the remaining unspecified value uniquely.

The real innovation here is the attempt to find the optimal measures of size, i.e., the Z_{1p}'s. For various reasons, such as the desire for a selfweighting sample, these measures are generally taken to be proportional to the total number of HU's or the total population in a county. In addition, it is not very often that a survey is designed for the sole purpose of estimating one or two parameters, as was the case here. It is of interest to note the result obtained for $Z_{1p}=N_{1p}/N$ (i.e., probability proportional to the number of HU's) in (34). In this case $V^2_{r_M}$ simplifies to:

$$V^2_{r_M} \doteq \frac{N}{tX^2} \sum_p^{T_1} \frac{(X_{1p}-RY_{1p})^2}{N_{1p}} + \frac{(1-R)\,Def_L Def_2}{\left(\frac{tW}{\bar{n}}\right)\bar{n}\,\bar{\bar{K}}\,R\,\pi} \quad (35)$$

where a negligible term has been discarded. As seen from (19), apart from the fpc, the second term in (35) is simply the relative variance for the familiar two-stage cluster sample of size $tW/\bar{n}$ clusters and tW HU's selected from across the entire State, without regard to the county from which they arise.

To determine the optimum combination of t, Z_{1p}, $\bar{n}$, and W, a cost function is needed.

B. Cost Function. A brief description of the cost equations will now be given. The cost model is similar to the earlier one except for an additional term to account for the variable cost associated with the sample PSU's. The cost function for a State is given by:

$$C_M = C_{M1}(t) + C_{M2}\left(t\frac{W}{\bar{n}}\right) + C_{M3}(tW) + C_{M4}\,t\sqrt{\frac{W}{\bar{n}}}\,\bar{A}, \quad (36)$$

where

C_M = total variable cost for the State,

C_{M1} = cost per sample PSU (includes the cost of hiring and supervising the interviewer in a sample PSU),

C_{M2} = cost per cluster,

C_{M3} = cost per HU,

C_{M4} = travel cost between clusters in the same PSU, and

$\bar{A}$ = average county land area (square miles) in the State.

The four cost coefficients were computed separately for each State.

C. Determining the Optima. The method by which the optima were approximated will now be described. As before, no exact mathematical solution exists, however, computer assisted iterative optimization solutions over all possible reasonable combinations of the unknowns ($t, Z_{1p}, \bar{n}, W$) are easily found. Separately for each State, we specified the following 360 combinations of the probabilities Z_{1p}, the workload W, and the cluster size $\bar{n}$:

1. Six sets of probabilities, Z_{1p}:

$$\frac{P_{1p}}{P}, \frac{K_{1p}}{K}, \frac{N_{1p}}{N}, \frac{Y_{1p}}{Y}, \frac{\sqrt{Y_{1p}K_{1p}}}{\sum_q^{T_1}\sqrt{Y_{1q}K_{1q}}}, \frac{1}{T_1}$$

where P_{1p} = total population in PSU p, and

$P = \sum_p^{T_1} P_{1p}$ = total population in the State.

2. Ten workloads, W: 50, 100, 150, 200, 250, 300, 400, 450, 500
3. Six cluster sizes, $\bar{n}$: 1, 2, 3, 4, 5, 6.

The second and third sets of the Z_{1p} listed above are slight variations of the conventional measure P_{1p} (the first set).

The fourth set was investigated because one would expect it to identify areas of large numbers of minority. It was also hoped that the fifth set of the Z_{1p} would identify "pockets" of high minority density. This set of basic draw probabilities is probably the most interesting of the six sets and the intuition behind it was based upon the sample allocation formula as given in III.C. Actually, it was completely unknown as to how well this fifth set would ultimately perform. The sixth and final set of the Z_{1p} was tested only for curiosity purposes and consistently resulted in ridiculous optima, as expected.

Separately, for each specific combination of the Z_{1p}, W, and $\bar{n}$, expression (34) was used to solve the integer number of sample PSU's, t, necessary to satisfy $V^2_{r_M} \leq .01$. The cost of each specific combination of possible optima was then determined by using (36). The following table shows the resulting minimum cost combinations and other information for each State but Arizona, which, as mentioned earlier, was unique in that 9 of its 14 counties were already county jurisdictions. As the table shows, the measures of size Y_{1p} and $\sqrt{Y_{1p}K_{1p}}$ both performed quite well. Not shown in the table is the fact that for a given State, whenever Y_{1p} was the optimum measure of size, then $\sqrt{Y_{1p}K_{1p}}$ was never far behind, and conversely. Except for Texas, all optima shown in the table were actually used in the sample selection. For Texas, three sets (rows) of optima are listed, with the first and second sets corresponding only to the Black or the Spanish minorities, respectively. The third combination listed was the one used in Texas and was approximately optimal when considering both the Black and Spanish minorities. As a matter of fact, in Texas, additional sets of Z_{1p} were investigated

which were functions of both the Black Y_{1p} and the Spanish Y_{1p}. However, these special measures of size generally performed worse than the conventional P_{1p}, which was ultimately used. Since the optima were fairly flat, it was not uncommon to find that $\bar{n}$=3,5, or 6 (along with the appropriate combination of Z_{1p}, W_2 and t) was approximately optimal, along with $\bar{n}$=4. In these toss-up cases, the set of optima with $\bar{n}$=4 was chosen because of the advantages of using established sampling procedures in our three frames. Finally, the last column in the table indicates the amount of money that was saved by our optimization procedure over an alternative procedure which fixes the $Z_{1p} = P_{1p}/P$ (the conventional measures) and then optimizes.

D. Stratifying the PSU's. After determining the above statewide optima, the first two restrictions imposed by our model (34) were relaxed. Each State was stratified using approximately equal size strata and one PSU was selected per stratum using the measures of size determined optimal for the State. The workloads were then slightly adjusted to reflect the differing stratum sizes. Strata were formed on the basis of the percent minority and the minority median family income in the counties. In addition, there was frequently one stratum in each State that contained counties with virtually no minorities. Although it is not desirable to have these small minority PSU's in sample, it was felt safer to guarantee one and only one such PSU in sample rather than take a chance of selecting none, one, or more than one.

Even though more than modest gains were expected from the stratification, the sample sizes were not reduced to reflect this gain. This decision was based upon the fact that there were considerable approximations both in developing (34) and in estimating the many county totals used in the optimization. The gains associated with the complete stratification have not been estimated, however, the gains associated with the inclusion of our certainty PSU's only, were expected to reduce the 10 percent CV to about 9.6 percent in each State. Under this modified model that considers our certainty PSU's, the between-PSU variance as a percent of the total variance ranges from 10 to 25 percent across the 8 States.

E. Within PSU Sampling. For the sample counties in each State the optimal cluster size was shown to be $\bar{n}$=4. Thus, within each of the three sampling frames in each county, cluster sizes of n=4 were employed. The workload in each sample county was allocated to the three frames exactly as described earlier for the county jurisdictions.

V. ALTERNATIVE SAMPLE DESIGNS AND THE 1978 VOTING RIGHTS SURVEY

This final section includes a brief discussion of the research into alternative sample designs that is currently taking place and of the upcoming 1978 Voting Rights Survey.

A. Alternative Sample Designs. The tendency for people to overreport voting and the resulting bias is a common problem in survey work. Although the 1976 sample design did not address this unfortunate phenomenon, it is planned to reduce the over-reporting bias, where possible, by ratio estimating to the actual overall number of votes cast as given by the jurisdictions themselves. In addition, the Bureau has begun research concerning two alternative sample designs that are expected to reduce the over-reporting bias at an affordable price.

The first alternative is a dual-frame sampling scheme. The two sampling frames in this scheme are (1) the usual Bureau frames described throughout this paper, and (2) county registration lists. A sample is drawn from each frame and a combined dual-frame estimator is employed. For a given amount of money, it is unknown as to whether or not the mean squared error of the dual-frame estimator is less than that of the conventional sample design estimator. Research is continuing in 12 county jurisdictions in an attempt to answer this question.

The second alternative sample design is a double sampling records check approach. In this design, the usual household survey is conducted and a subsample of the surveyed households is then selected. The voting responses for the persons in these subsampled households are then checked against voter and registration lists and an estimator reflecting the observed over-reporting in the subsample is formed. Again, our research seeks to determine the cost effectiveness of this double sampling scheme.

B. 1978 Voting Rights Survey. The Bureau conducted the 1976 Voting Rights survey in the 93 jurisdictions and the research discussed above for about $5,000,000. The 1976 survey, however, is small in both price and the number of covered jurisdictions compared to the 1978 Voting Rights Survey currently being planned. For the 1978 survey, the Bureau has been directed to treat each individual county in the nine States as a jurisdiction in its own right. In addition, the town and county jurisdictions covered in the 1976 survey are to be retained in 1978. Thus, the Bureau is expected to be given about $40,000,000 to conduct sample surveys or censuses in about 950 town and county jurisdictions in November 1978.

The innovative sample design strategy presently being planned for the 1978 survey is highly analytic in nature. We are attempting to divorce ourselves from the relatively artificial 10 percent CV reliability requirement concept and design the survey with the analyst and decision maker in mind. The power function is the key concept in our unique design. As of this writing, it is felt one of the best ways to spend the $40,000,000 is to design the 1978 survey so that in each sample jurisdiction, the probability of concluding the White voter participation rate is more than 3 percentage points higher than the minority voter participation rate (versus concluding the difference is exactly 3 percentage points), is equal to .10 when the true difference is 3 percentage points (a type I error), and is equal to .90 when the actual difference is 10 percentage points (a correct conclusion). In addition, the budget for the 1978 survey includes funds for a 100 percent voting records check, thus eliminating the over-reporting bias.

TABLE OF STATE OPTIMA

State	Minority of interest	π	R	Z_{1p} proportional to:	W	$\bar{n}$	t	Total sample size =tW	Total 4/ sample size if statewide SRS	Dollars 5/ saved over $Z_{1p} = \frac{P_{1p}}{P}$
Alaska	Native Alaskan	.13	.43	$\sqrt{Y_{1p}K_{1p}}$	150	4	11	1650	1050	$20,000
Alabama	Black	.23	.43	Y_{1p}	100	4	11	1100	600	$ 7,000
Georgia	Black	.22	.38	$\sqrt{Y_{1p}K_{1p}}$	100	4	13	1300	750	$ 8,000
Louisiana	Black	.26	.43	Y_{1p}	100	4	9	900	500	$ 3,000
Mississippi	Black	.31	.43	Y_{1p}	100	4	9	900	450	$ 2,000
So. Carolina	Black	.26	.39	Y_{1p}	100	4	11	1100	600	$ 2,000
Texas	Black 1/	.11	.43	$\sqrt{Y_{1p}K_{1p}}$	150	4	12	1800	1250	$15,000
	Spanish Heritage 2/	.14	.43	$\sqrt{Y_{1p}K_{1p}}$	100	4	13	1300	1050	$20,000
	Black, Spanish Heritage 3/	-	-	P_{1p}	100	4	21	2100	1250	0
Virginia	Black	.16	.43	$\sqrt{Y_{1p}K_{1p}}$	100	4	12	1200	800	$11,000

1/ Considers Black only, ignores Spanish Heritage. This design for Black would yield an unacceptable 24 percent CV for Spanish Heritage.
2/ Considers Spanish Heritage only, ignores Blacks.
3/ This design jointly yields a 10 percent CV for Blacks and a 9.2 percent CV for Spanish.
4/ For comparison with tW, this column gives the sample size for a statewide simple random sample (SRS)
5/ This column gives the savings over the conventional design using probability proportional to total population, i.e., $Z_{1p} = P_{1p}/_{P}$.

REMINDER

π = fraction minority of interest
R = minority voting rate
Z_{1p} = single draw probabilities
W = Within PSU sample size
$\bar{n}$ = average HU cluster size
t = number sample PSU's

REFERENCES

1. W. G. Cochran. Sampling Techniques. 2nd ed. New York: Wiley and Sons, 1963.
2. J. Durbin. Some Results in Sampling When the Units are Selected with Unequal Probabilities. Journal of the Royal Statistical Society, Series B, (1953), Vol. 15, pp. 262-269.
3. M. H. Hansen, W. N. Hurwitz and W. G. Madow. Sample Survey Methods and Theory, Vol. 1, Methods and Applications. 1st ed. New York: Wiley and Sons, 1953.
4. R. Scammon. American Votes 10. 10th ed. Washington, D. C.: Congressional Quarterly, 1972.
5. M. Thompson and G. Shapiro. The Current Population Survey: An Overview. Annals of Economic and Social Measurement, (1973), Vol. 2, No. 2.

ACKNOWLEDGMENTS

The author would like to thank Bob Jewett and Duane Hybertson for their excellent computer programming and Edith Oechsler for her careful typing; all of the U. S. Census Bureau.