# SYSTEMATIC SELECTION OF PPS SAMPLES FOR SEVERAL SURVEYS FROM THE SAME UNIVERSE

Julia L. Bienias, David L. Hubble, and Charles H. Alexander
U.S. Bureau of the Census, FOB 3, Room 3725, Washington, DC 20233

KEY WORDS: Systematic sampling, multiple sampling, PPS sampling

## I. BACKGROUND

In the redesign and reselection of sample for the Census Bureau's ongoing household surveys, the same sampling frame is used for six surveys, the Current Population Survey (CPS), the American Housing Survey (AHS), the National Crime Survey (NCS), the National Health Interview Survey (HIS), the Survey of Income and Program Participation (SIPP), and the Consumer Expenditure set of surveys (CE).

These surveys use multiple frames. Each geographic area in the country is sampled under one of the following procedures: 1) census address lists supplemented by a frame of building permits for units constructed since the census, in areas where the census addresses are easily locatable and building permits are required; 2) an area frame in which units are listed at the time of interviewing, for units built before the census supplemented by building permits for later units, in areas where the census addresses are incomplete but building permits are required; 3) an exclusively area frame in areas where building permits are not required (see U. S. Bureau of the Census, 1978). The permit frame is unduplicated from the others by determining the year in which a sample structure was built. Starting with the 1980 redesign, HIS does not use the first procedure, replacing it with the second procedure instead.

In the redesign following the 1990 census, the basic secondary sampling unit will be the census block. In previous redesigns, the larger enumeration district (ED) was the secondary unit, because blocks were defined for only part of the nation. In the 1990 redesign, blocks will be sorted first, and then a systematic sample of clusters of housing units (typically an expected four units) will be chosen. Under this procedure, the probability that a block will have a cluster chosen from it is proportional to its size. In previous redesigns, several surveys could use the same secondary unit, although the surveys were not allowed to select the same households.

Allowing several surveys to use the same block or ED causes some operational problems. One problem is that listings of blocks in the area frame need to be cross-referenced. The first survey to use a block in the area frame must have a field representative list the households in the block; the sample units are selected from this list. Later surveys using this block must obtain this list, update it, and select their sample households from the unused part of the updated list, to eliminate the chance of selecting the same unit for two surveys. This cross-referencing requires a complex record-keeping system to keep track of who has listed or updated any given block, and it means that sample reductions by any survey interferes with the operation of other surveys. Even more complex cross-referencing is required when field representatives obtain lists of building permits for the permit frame. This complexity is a disadvantage of sharing the same block, although the cost savings of being able to share the lists are a compensating advantage.

Several of the surveys have special features which make it particularly troublesome for them to share blocks with other surveys. For example, the HIS is collected under different authorizing legislation than the other Census Bureau surveys, so confidentiality restrictions prevent HIS from using listings made for the other surveys. In the 1980 design, HIS was kept away from other surveys' EDs by selecting all surveys by a systematic sampling procedure (using a cumulative measure of size) with the same sort order for the EDs; HIS sample was located halfway between the "hits" for the other surveys. In the few areas where EDs were shared, special clerical operations were needed to try to eliminate the duplicate selection of addresses.

Requiring all the surveys to use the same sort order of EDs restricts the ability of the surveys to reduce variance by sorting on characteristics that are related to the variables of interest to the survey. Because of this, CE and AHS select individual addresses from the census address lists using their own sort order of units; these surveys do not use block or ED as a secondary unit in the address frame. Because their sample units may end up being in blocks which are in sample for the other surveys, special operations are needed to eliminate the duplication. These surveys initially select more units than they need and then eliminate those that have also been selected for the other surveys.

Because of these various operational problems which arise when several surveys share the same block, our research for the 1990s sample redesign is considering several schemes for giving some or all of the surveys a "subuniverse" of blocks for their exclusive

use. This idea could be applied to all six surveys, but the most likely applications would be to give HIS and possibly AHS their own sets of secondary sampling units, possibly just in the area frame.

The various proposed subuniverse schemes need to be evaluated based on their effects on operational complexity and associated system design costs, ongoing field and clerical costs, and variance. This paper concentrates on the effect on variance of one of the schemes for reserving entire blocks for a single survey.

The obvious way to reserve blocks for exclusive use by one survey is to have that survey select its sample prior to the others and to remove its sample blocks from the universe before the other surveys select their samples. However, blocks are removed with probability proportional to their sizes. This means that the first survey will remove proportionally more large blocks than small ones. When it comes time for the second survey to select its sample, the universe that is "left over" will under-represent large blocks. We initially sought to find a simple factor to adjust for this (based on which units were selected for the first survey), but we were unable to find one.

An alternative approach is to select the sample in two stages. Blocks are partitioned randomly into "subuniverses"; each block is given equal probability of being in the subuniverse. In the first stage, the first survey chooses one of the subuniverses using equal probability selection. This subset of blocks is then reserved for the survey's exclusive use. Next, the sampling fraction from the subuniverse is increased so that the expected sample size is the same as it would be with the usual approach. For example, if the subuniverse contained one-tenth of the blocks, the sampling interval would be one-tenth of that used for the entire PSU. The sample blocks for the first survey are then selected from the subuniverse with probability proportional to size. The universe of blocks left for the remaining surveys will be a representative sample of the original universe of blocks.

Each subuniverse can be created as a simple random sample of blocks from the original universe, as a stratified sample with equal sampling fractions in the strata, or as a systematic sample after the blocks have been sorted in some appropriate order. In this paper, we will assume that the systematic approach is used.

The subuniverse approach ensures that the first survey will have exclusive use of its sample blocks, avoiding the operational problems described in the previous section.

The major uncertainty about the subuniverse approach is its potential for increasing sampling variance, compared to the usual systematic selection of clusters, in which blocks are "hit" with probability proportional to size from the full universe. There are several reasons why the variance might be expected to increase, two of which are described briefly here in this abbreviated report.

It may not be possible to take full advantage of an optimal sort of blocks. If the blocks are sorted according to characteristics that are highly correlated with the variables of interest to the survey, then a systematic sample from the entire universe using the cumulative measure of size can improve the variance compared to sampling with less control over the selection. Selecting a subuniverse first by assigning blocks to subuniverses with equal probability regardless of their measures of size reduces this control, which may increase variance.

The subuniverse method may increase variability in the number of sample units chosen from each PSU. If a given subuniverse includes larger-than-average blocks, and the sampling fraction has been adjusted based on the expected size of a subuniverse, then the sample size from that subuniverse will exceed the expected size, and vice versa if the blocks in the subuniverse are smaller than average. This variability is not present with the usual method of selection.

This second problem can be eliminated by modifying the method by a) partitioning the universe within each PSU into subuniverses with approximately equal numbers of housing units each, b) choosing a subuniverse with probability proportional to total subuniverse size, and c) varying the within-subuniverse sampling rate so the total number of hits is constant. This modification still approximates PPS selection of blocks, but it adds some additional complexities, especially for subsequent surveys which select units from the remaining part of the universe after the first survey has been selected.

In the current research, we sought to evaluate this possible solution to the second problem above. This evaluation was accomplished through simulating the subuniverse sampling procedure, using data on the distribution of block sizes from PSUs included in the 1990 test censuses and using a range of generally realistic assumptions about the distribution of block means for the characteristic of interest (taken to be household income). In the results presented here we assume that the block's mean income is independent of the block size. The within-block component of

variance is totally omitted from the analysis by assigning the block mean to each household in the block; this seems reasonable, since the method of sorting and selecting blocks probably has little effect on the within-block variance.

## II. METHOD

### A. Block Distributions

To measure the effect of creating subuniverses on the variance of an estimator, we used different frequency distributions of block sizes from actual PSUs included in the U. S. Census 1988 Dress Rehearsal. Sites chosen for the Dress Rehearsal were to include, for the most part, contiguous whole cities or counties that would require the implementation of the various enumeration methods and procedures needed for the 1990 Census.

We chose one area from each of the four distinct types of areas in which the Dress Rehearsal was conducted. These were, from highest to lowest population density: a hard-to-enumerate large city (St. Louis), a small city (Columbia, in Boone County, MO), a suburban/ rural area (Cooper-Morgan Counties, MO), and a very sparsely populated rural area (Grant County, WA) (see Note). To simplify the programming of the simulation (described below), we randomly deleted up to 7 blocks from these areas to create an integral multiple of 10 blocks.

### B. Characteristic of Interest

We chose to calculate the variance of estimated mean household income because it is a commonly measured variable which is of direct interest to several of the Census Bureau's surveys, and it is highly correlated with other variables of interest and can therefore be used to sort units to create approximate stratification when using systematic sampling. For the current research, we chose to impute the block mean values for household income. In later research we will use actual data. To impute household income data we used an estimated household income mean of 31,010 (s.d. = 23,141) for 1986, the latest year for which data were readily available (U.S. Bureau of the Census, 1987).

Blocks of size one are simply single housing units, and values for income were randomly chosen from a lognormal distribution with the mean and variance given above, as such a distribution has been shown to fit income data reasonably well.

In imputing values for a block of size $b > 1$, we are choosing values of sample means based on samples of size $b$. Such values can be reasonably chosen from the sampling distribution of the mean of samples of size $b$ from a lognormal distribution. If we consider the units in the block as though they are random samples from the population, then by the Central Limit Theorem, the sampling distribution of block means approaches a normal distribution for large $b$. That is, the mean value for a block could be chosen from the distribution N(31010, (23141)^2/b). To simplify the calculations, a normal distribution was used for all block sizes greater than one, even though that is not "large" $b$. However, the means of blocks are not simply means of random samples of size $b$ drawn from the PSU. Instead, they can be considered as the means of units clustered in groups of size $b$. To take this into account, a within-block intraclass correlation, denoted rho, was incorporated into the variance, drawing random values from a normal distribution with mean 31,010 and variance (23141)^2(1+(b-1)rho)/b. These imputed values were constrained to be positive.

Using different random seeds, five data sets were created for each block distribution for each of the two levels of rho (0 and .20).

### C. Simulation

Five sets of simulations were conducted choosing 2, 5, and 10 subuniverses, which are possible numbers of subuniverses that might be created for the Census Bureau's surveys. The blocks were sorted in ascending order by size, and systematic samples of blocks were formed to create the specified number of subuniverses. For each data file of block size and imputed income, sampling variances were calculated for each of the following designs by generating all possible samples. (See the Appendix for the formulae and their derivations.) The calculations used routines developed in VAX C. For consistency, a sample size of 100, which is a typical interviewer workload for HIS, was drawn from each geographic area. Although the actual surveys draw clusters of households, for simplicity a single housing unit was the ultimate sampling unit.

#### Design 1: All Blocks

Draw $K$ systematic samples of approximately size $m$ from across the entire PSU. The estimator is a simple sample mean which is unbiased if all the possible samples are of exactly size $m$.

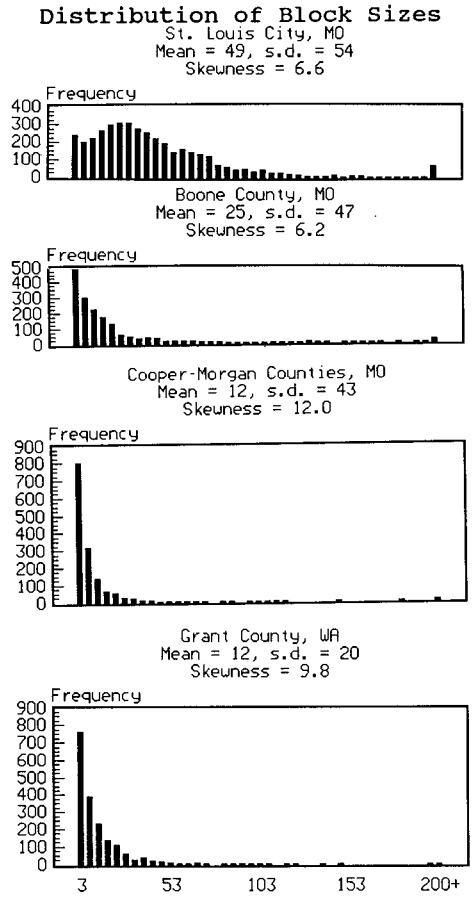
#### Design 2: PPS Subuniverses

Choose one of the $S$ subuniverses with probability proportional to its number of housing units and then choose one of $K$ systematic samples of approximately size $m$ from it. The estimator is a simple sample mean which is unbiased if all the possible samples are of exactly size $m$.

## III. RESULTS

The mean number of housing units per block and the distribution of blocks sizes did indeed vary among the four areas chosen. The average block size

and skewness of the distributions are provided with the graphs of the distributions below.




Number of Housing Units per Block
(scale is midpoints of groups of 5)

Table 1 presents the average number of housing units in each subuniverse under the different designs, by the four geographic areas. The degree of variability in subuniverse size is small and remains relatively constant across the different designs, for each area. Thus the effect on variance of variation in subuniverse size would not seem to be dramatic, especially for the larger areas.

For each of the four geographic areas and designs, the average of the standard deviations of the estimators of mean household income was computed, averaged over five replications (five sets of imputed income data). These averages and their standard deviations are presented in Table 2 for the two levels of rho.

For rho = 0, it can be seen that there is no clear pattern of variance increase or decrease for the subuniverse designs. Whereas it was expected that the variance should increase monotonically as one moves from drawing samples from across the PSU to drawing them from within an increasing number of subuniverses, in St. Louis, the large city, the variance for the subuniverse designs was actually less than that for the samples drawn across the entire PSU.

However, it should be noted that treating blocks as independent samples of units (considering the units within the block to be heterogeneous) probably artificially decreases the variance. Under this approach, the variance of the sampling distribution is the variance of household income deflated by the size of the block. To create a more realistic situation of greater homogeneity of units within blocks, a rho of .20 was used, which adds an inflating factor to the sampling distribution variance.

For rho = .20, the results are different. First, it should be noted that all of the standard deviations in Table 3 are larger than the corresponding ones in Table 2. This is because the variance of the hypothesized income distribution increases as the value of rho increases. The impact of this is greatest with large blocks, for which the inflation term on the variance of the distribution of means of samples the size of the block is greatest.

A higher value of rho affects the impact of the subuniverse designs. Because increasing rho adds more variability to the distributions of the means of large blocks, those areas with highly skewed block distributions suffer a greater impact when subuniverses are used. The few large blocks are not distributed equally (e.g., there may only be three such blocks and 10 subuniverses), and thus from replication to replication, the values of the estimators vary widely as the large blocks have widely varying means.

Cooper-Morgan PSU, the suburban/rural area, seems to suffer the worst variance increment, and its distribution is most skewed. No matter how few subuniverses are created, they do not share the large blocks equally, because there is a wide spread in blocks sizes among the very largest blocks. For example, the difference between the number of housing units in the largest block and in the fifth largest block is 459 in Cooper-Morgan PSU, and ranges between 260 and 285 in the other three areas.

## IV. DISCUSSION

At this preliminary stage, the results seem to indicate that whether a subuniverse design leads to an increment in variance depends on the assumed rho, especially as it applies to large blocks. However, as only four areas were examined, more research is required before a conclusion can be reached.

The problem with "large" blocks is a

potential drawback of the subuniverse method. Although there are relatively few such blocks, depending on how "large" is defined, they contain a disproportionate amount of the population. If they have different characteristics than other blocks, the increased variance due to these blocks can be substantial.

One solution may be to sample these blocks separately from the others, allowing them to be shared by several surveys. Alternatively, the blocks can be subdivided a priori and the subblocks treated as full blocks in creating the subuniverses, if the geography of the large blocks is known. Having to develop special procedures for either of these methods reduces the basic operational simplicity of the subuniverse approach.

Before decisions can be made regarding the feasibility of creating subuniverses, the effect on the bias and variance for subsequent surveys needs to be researched. The estimator presented here is unbiased, but using a PPS selection of subuniverses affects the "left over" universe. An SRS selection of subuniverses and a slightly biased estimator might be preferable.

In addition, different distributions of the characteristic of interest and different correlations of the characteristic of interest with block size should be considered. We intend to examine actual data for different characteristics of interest, both to understand the effect on variance in a real application and to be able to test data that are correlated with block size.

## V. ACKNOWLEDGMENTS

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. The authors wish to thank Patrick Cantwell, Deborah Fenstermaker, Brenda Kelly, Gary Kamer, and Elizabeth Sweet for their assistance. For a longer version of this report, please write to the authors.

## VI. NOTE

More information on the areas included in this study is available from the authors.

## VII. REFERENCES

Cochran, W.G. (1977). Sampling Techniques (3rd ed.). New York: Wiley.

U.S. Bureau of the Census. (1978, January). The Current Population Survey: Design and Methodology. Washington, D.C.: Government Printing Office.

U.S. Bureau of the Census. (1987, December). Statistical Abstract of the United States: 1988 (108th ed.). Washington, D.C.: Government Printing Office.

Table 1. Average Number of Housing Units Per Subuniverse in Each Design (Standard Deviations in Parentheses), by Geographic Area

| | Number of Subuniverses[1] | | |
|---|---|---|---|
| | 2 | 5 | 10 |
| Large City | 98,666 (296) | 39,466 (274) | 19,733 (284) |
| Small City | 21,554 (149) | 8,622 (231) | 4,311 (213) |
| Sub./ Rural | 9,015 (267) | 3,606 (275) | 1,803 (245) |
| Rural | 11,384 (77) | 4,553 (144) | 2,277 (138) |

Table 2. Average Standard Deviations of Estimators of Mean Household Income, over 5 Replications, for Within-Block Intraclass Correlation of 0 and .20 (Standard Deviations in Parentheses), by Geographic Area

| | Number of Subuniverses[1] | | | |
|---|---|---|---|---|
| | None (whole PSU) | 2 | 5 | 10 |
| A. Rho=0 | | | | |
| Large City | 340 (10) | 333 (22) | 320 (27) | 319 (24) |
| Small City | 428 (26) | 446 (42) | 435 (19) | 485 (35) |
| Sub./ Rural | 626 (60) | 637 (69) | 665 (54) | 686 (75) |
| Rural | 604 (61) | 579 (42) | 595 (47) | 652 (98) |
| B. Rho = .20 | | | | |
| Large City | 977 (88) | 979 (84) | 1003 (125) | 1087 (109) |
| Small City | 1001 (74) | 1136 (276) | 1435 (83) | 1948 (333) |
| Sub./ Rural | 958 (102) | 1306 (318) | 2101 (601) | 2879 (950) |
| Rural | 989 (132) | 1014 (102) | 1255 (179) | 1470 (214) |

[1]See text for a description of the designs used.

## Appendix

Following are the estimators and the derivations of their variances for the two sample designs. To simplify the notation and derivation, the sample size is denote by $\underline{m}$ for all samples, although sample size can vary from m - 1 to m + 1. In the actual variance computation the exact sample size was used.

Notation

$m \equiv$ sample size

$N \equiv$ number of subuniverses

$K \equiv$ number of systematic samples across the entire PSU

$M_s \equiv$ number of housing units in subuniverse s

$M_o = \sum_{s=1}^{N} M_s \equiv$ total number of housing units in the whole PSU

$K_s \equiv$ number of systematic samples in subuniverse s

Subscripted $\underline{y}$ denotes the value of the characteristic of interest (here, income) for the subscript cell

$Y \equiv$ total value of the characteristic over all housing units in the PSU

$Y_s \equiv$ Total value of the characteristic over all housing units in subuniverse s

$\bar{Y} = \frac{Y}{N} \equiv$ mean total income per subuniverse

$Y = \frac{Y}{M_o} \equiv$ true mean income per HU (the value being estimated)

Design 1: All Blocks

This variance is simply the variance of a simple mean from a systematic sample, an estimator which is unbiased for samples of size exactly $\underline{m}$ (Cochran, 1977, pp. 207-208).

$$\hat{\bar{Y}}_1 = \bar{y}_{1k} = \frac{1}{m}\sum_{j=1}^{m} y_{kj} \text{ for the } k^{th} \text{ systematic sample}$$

$$\hat{V}(\bar{Y}_1) = \frac{1}{K}\sum_{i=1}^{K}\left[\bar{y}_{1i} - \bar{Y}\right]^2$$

Design 2: PPS Subuniverses

$$\left(P(\text{subuniverse } s) = \frac{M_s}{M_o}\right)$$

$$\hat{\bar{Y}}_2 = \bar{y}_{2sk} = \bar{y}_{1sk} = \frac{1}{m}\sum_{j=1}^{m} y_{skj}$$

for the $k^{th}$ systematic sample in subuniverse s (see Cochran (1977, p. 294).

$$E(\bar{Y}_2) = E_1(E_2(\bar{y}_{1sk}|s)) = E_1(\bar{Y}_s)$$

because $(\bar{y}_{1sk}|s)$ is simply a sample mean for a systematic sample and it is unbiased for the true mean within that subuniverse.

For $mK_s = M_s$, we have

$$E(\hat{\bar{Y}}_2) = \sum_{s=1}^{N}\frac{1}{N}\bar{Y}_s = \frac{1}{M_o}(Y) = \bar{Y}$$

So, $\hat{\bar{Y}}_2$ is an unbiased estimator of $\bar{Y}$ for samples of size exactly $\underline{m}$.

Following the technique of Cochran (1977, p. 276), we have:

$$V(\hat{\bar{Y}}_2) = V_1\left[E_2\left(\bar{y}_{1sk}\middle|s\right)\right] + E_1\left[V_2\left(\bar{y}_{1sk}\middle|s\right)\right]$$

$$= V_1(\bar{Y}_s) + E_1\left[\frac{1}{K_s}\sum_{i=1}^{K_s}(y_{1si} - \bar{Y}_s)^2\right]$$

$$\text{But } V_1(\bar{Y}_s) = \sum_{s=1}^{N}\frac{M_s}{M_o}(\bar{Y}_s - \bar{Y})^2$$

(Cochran, 1977, p. 252)

$$\text{So, } V(\hat{\bar{Y}}_2) = \sum_{s=1}^{N}\frac{M_s}{M_o}(\bar{Y}_s - \bar{Y})^2 + \sum_{s=1}^{N}\frac{M_s}{M_o}\left[\frac{1}{K_s}\sum_{i=1}^{K_s}(y_{1si} - \bar{Y}_s)^2\right]$$

But, $\sum_{s=1}^{N}\frac{M_s}{M_o}\bar{Y}_s = \bar{Y}$, and therefore,

$$\sum_{s=1}^{N}\frac{M_s}{M_o}(\bar{Y}_s - \bar{Y})^2 = \sum_{s=1}^{N}\frac{M_s}{M_o}\bar{Y}_s^2 - \bar{Y}^2$$

$$\text{So, } \hat{V}(\bar{Y}_2) = \frac{1}{M_o}\left[\sum_{s=1}^{N}\sum_{i=1}^{K_s}\frac{M_s}{K_s}y_{1si}^2\right] - \bar{Y}^2$$