USING SAMPLE INFORMATION FOR STRATIFICATION

David R. Judkins, Hertz Huang and Gary M. Shapiro, U.S. Bureau of the Census

I. INTRODUCTION

There are situations where one wants to use a subset of sample primary sample units (PSUs) for a survey. This paper compares two strategies for accomplishing this. Consider, for example, taking a subsample of the Current Population Survey (CPS), conducted by the Census Bureau for the Bureau of Labor Statistics to collect labor force data. This survey is a two-stage stratified design of 729 sample areas, one sample PSU per stratum. We might be interested in a subsample of CPS PSUs for a much smaller scale one-time survey. For such a survey, it would be very inefficient to conduct it in all the CPS sample PSUs. It would, of course, be possible to independently stratify and select sample PSUs for the new survey, but that would cause such high sampling and interviewing costs in PSUs not in the CPS sample as to usually be infeasible. Also, extra calendar time, not usually available, is needed to independently stratify PSUs. Thus, a preferred design would guarantee complete overlap in sample PSUs between the two surveys. Throughout this paper, we assume that complete overlap is a requirement for the sample design of the new survey.

Some aspects of the general problem of how best to take a subset of sample PSUs for a reduced survey are addressed in this paper. We confine our remarks to the case where the original survey selected one PSU per stratum, possibly in a dependent manner between strata. One general approach is to collapse together strata from the original survey. For example, if about a 1/2 subsample of PSUs is desired, one might form pairs of the nonself-representing strata and retain only one sample PSU from the pair with probability proportionate to the size of stratum. Collapsing would be done by using available information for auxiliary variables believed or known to be correlated with the most important survey characteristics to form homogeneous collapsed strata, or superstrata. Other approaches for taking the subsample could also be used: sample PSUs could be sorted in an appropriate manner and a systematic sample taken.

The issue addressed in this paper is which of two general strategies for subsampling a set of sample PSUs is better: use estimates of the auxiliary variables based on the original survey's sample PSUs or each stratum as a whole. The first strategy uses information on the outcome of PSU selection by the original survey in determining the sample design; the second does not. We call them the informed and the uninformed strategies. At first glance, the informed strategy appears potentially biased. To many people, it appears to yield higher variances than use of full strata data. However, we show in this paper that use of sample PSU information is unbiased. We also explain why it tends to yield lower between-PSU variances for the stratification variables or function. Section II of the paper gives a non-mathematical explanation for why use of sample PSU data is preferable. Section III compares the components of variance between the two strategies and gives an example to further clarify the comparison. Section IV provides variance estimators for both one sample PSU and two sample PSUs per superstratum in the new survey. The methodology for two sample PSUs is innovative and of potential application to other situations.

II. ARGUMENT FOR USE OF SAMPLE DATA

The issue to be discussed here is whether the use of data on the whole strata or on the sample PSUs in designing a new survey tends to yield the smaller true between-PSU variance for the new survey. (Estimating the between PSU variance is addressed in Section IV.) It is helpful to think of this as a double sampling situation. In the first phase of sampling, we stratify the PSUs for the original survey and select one sample PSU per stratum. Thus, the initial sample consists of the original survey sample PSUs. If the new survey is considerably smaller than the original survey, the natural method of subsampling these sample PSUs is superstratification. (We will explicitly refer only to this specific method for the rest of this section. The discussion is, however, equally relevant to other methods for subsampling.) In the second phase then, we stratify these sample PSUs (or equivalently, the strata represented by the sample PSUs) into superstrata and select one (or more) of the sample PSUs per superstratum. Viewed in this way, if we formed the superstrata based on characteristics of the whole strata, we would be failing to use the information on the outcome of PSU selection obtained in the first phase of sampling. (This information is used directly in the informed strategy when we make strata homogeneous with respect to the original-survey sample PSUs.) Thus, use of the stratum data rather than the sample PSU data utilizes less, not more, information. Since the goal in the second phase of sampling is to obtain a set of sample PSUs as much like the original set of sample PSUs as possible, the non-use of this information would be expected to increase variances.

This can also be looked at in a slightly different manner. There are three components of variance: the within-PSU, the between-PSU within-stratum variance, and the between-stratum (within superstratum) variance. The within-PSU-variance is not predictably affected by the choice of strategy. The between-PSU-within-stratum variance is fixed once the original survey is designed and unaffected by whether we use stratum or sample PSU characteristics for subsampling. The between-stratum variance that is relevant here is an expected conditional variance, conditioned on the original-survey sample PSUs. We consider this component more carefully. If we were not constrained to select only original-survey sample PSUs, then we would of course want to ignore sample-PSU characteristics. However, we should expect use of sample-PSU characteristics to be useful in minimizing the expected conditional variance under consideration since it is unreasonable to expect to minimize the expected conditional variance by ignoring what we know about the conditions, namely, the identity and characteristics of the sample PSUs. Granted, there can exist situations where use of known conditions may not reduce an expected conditional variance, but it is difficult to conceive of a situation in which use of known conditions would increase the expected value of a conditional variance.

There are two side comments of interest: (1) If the stratification criteria are the same for the original and new surveys, then it makes little or no difference which of the two strategies is used, because if two sample PSUs were similar to each other, so must be the strata that they represent. (2) If one is designing the smaller survey at the same time as the larger survey, it is still preferable to select the larger survey's PSUs first so that a double sampling approach can be used. Only if the smaller survey need not use a subset of the larger survey's sample PSUs, will it be better to independently stratify the smaller survey to take full advantage of the survey's stratification criteria.

III. VARIANCE DERIVATION WITH EXAMPLE

This section contains a technical comparison of the two strategies. We first make formal definitions of the strategies. We then define a very general estimator of population totals and show that it is unbiased under either strategy. Using a decomposition of the variance of this estimator, we argue that the most reasonable approach to the problem of minimizing the total variance on characteristics of interest is to minimize the between-strata variance of ancillary characteristics. We close the section with a concrete, though artificial, example where the total variance is smaller with the informed strategy.

Strategy Definition

Let D be the set of all PSUs. As stated previously, we are assuming that the original survey drew one PSU per stratum with dependence between strata. Thus, the only subsets of D that are admissible sets of sample PSUs for the original survey are those that consist of exactly one PSU from every stratum. Let G be this set of all sets of PSUs that were admissible for the original survey.

For every $g \epsilon G$, let H_g be the set of all subsets of g. There is an important correspondence between the elements of H_g and $H_{g'}$ for every g and $g' \epsilon G$.

Let $g = \left\{ d_1, d_2, \ldots, d_m \right\}$ and $g' = \left\{ d'_1, d'_2, \ldots, d'_m \right\}$. A typical element of H_g is $h = \left\{ d_1, d_2 \right\}$. The corresponding element of $H_{g'}$ is $h' = \left\{ d'_1, d'_2 \right\}$; i.e., the subset of g' that contains one sample PSU from each of the strata represented by the sample PSUs in h. We call this correspondence $\pi_{gg'}$ and write $h' = \pi_{gg'}(h)$ for $h \epsilon H_g$.

In order to select $g \epsilon G$, the original survey defined some probability measure μ on G. This measure may be arbitrary except that every PSU must have a nonzero chance of selection. A strategy for selecting a subset of g for the new survey is just a method for defining a probability measure on H_g. Given the uncountable number of possible measures on H_g, it is clear that some algorithm is needed. We do not discuss specific algorithms. Rather, we are concerned with what constraints are placed on whatever algorithm is actually used.

Under the informed strategy, any available information on the PSUs contained in g may be used to define the measure $\lambda(g)$ on H_g. For example, if unemployment is related to characteristics of interest and it is known that some of the PSUs selected for the original survey have zero unemployment while the rest have 100 percent unemployment, then $\lambda(g)$ should give positive measure only to those $h \epsilon H_g$ that contain a balance of high and low unemployment PSUs. Note that $[\lambda(g)](h)$ is the probability that the new survey selects $h \epsilon H_g$ given that the original survey selected g.

The uninformed strategy, on the other hand, spurns this information. It places a restriction on the algorithm of choice to only consider those measures $\nu(g)$ on H_g such that

$$[\nu(g)](h) = [\nu(g')](\pi_{gg'}(h)) \ \forall \ h \epsilon H_g \text{ and } \forall \ g,g' \epsilon G. \qquad (1)$$

In other words, this means that the probability of selecting h for the new survey given that g was selected for the original survey is equal to the probability of selecting the natural correspondent of h given that g' was selected for the original survey. In a sense then, $[\nu(g)](h)$ is the probability that the strata represented by the sample PSUs in h will be selected for the new survey. To put it yet another way, the probability of h given g depends only on the characteristics of the entire strata represented by the sample PSUs in H_g. Given this constraint on $\nu(g)$, we simply write ν for $\nu(g)$.

To summarize the notation developed so far:

- D = All PSUs,
- G = All possible sets of original-survey sample PSUs,
- μ = Original survey measure on G,
- H_g = All possible sets of new-survey sample PSUs given $g \epsilon G$
- $\lambda(g)$ = Informed strategy measure on H_g,
- ν = uninformed strategy measure on H_g

Estimator Definition

Let Y_d be the count of units (persons, households, etcetera) with some characteristic for $d \epsilon D$. The quantity to be estimated is Y, the sum of Y_d over $d \epsilon D$. We will assume that the within-PSU sampling is independent from PSU to PSU and is independent of the selection of sample PSUs. Let $\hat{Y}_d$ then be some unbiased estimator of Y_d. We next define binary functions that indicate whether a PSU is selected for the two surveys.

Let $\delta(d,g) = \begin{cases} 1 & \text{if } d \epsilon g, \\ 0 & \text{otherwise, and} \end{cases}$

$\beta(d,h) = \begin{cases} 1 & \text{if } d \epsilon h, \\ 0 & \text{otherwise.} \end{cases}$

Then the estimator of Y that we discuss is

$$\hat{Y} = \sum_{d\in D} \frac{\delta(d,g)\beta(d,h)\ \hat{Y}d}{E_\mu\delta(d,g)\ E_{\lambda(g)}\beta(d,h)}$$ for the selected $g\in G$ and the selected $h\in H_g$.

Note that $E_\mu\delta(d,g)$ is the probability of selecting PSU d for the original survey and that $E_{\lambda(g)}\beta(d,h)$ is the conditional probability of selecting PSU d for the new survey given that the original survey selected PSU set g.

Also note that we must have $E_\mu\delta(d,g)>0\ \forall\ d\in D$ and $E_{\lambda(g)}\beta(d,h)>0\ \forall\ d\in D$ and $\forall g\in G$ such that $\mu(g)>0$ and $\delta(d,g) = 1$.

Proof of Unbiasedness of Estimator

Since the within-PSU sampling is independent of the PSU sampling and $\frac{\delta(d,g)}{E_{\lambda(g)}\beta(d,h)}$ is fixed given g, we have that

$$E\hat{Y} = \sum_{d\in D} \frac{Y_d}{E_\mu\delta(d,g)}\ E\left(\frac{\delta(d,g)\beta(d,h)}{E_{\lambda(g)}\beta(d,h)}\right)$$

$$= \sum_{d\in D} \frac{Y_d}{E_\mu\delta(d,g)}\ E_\mu\left[E_{\lambda(g)}\frac{\delta(d,g)\beta(d,h)}{E_{\lambda(g)}\beta(d,h)}\,\Big|\ g\right]$$

$$= \sum_{d\in D} \frac{Y_d}{E_\mu\delta(d,g)}\ E_\mu\delta(d,g)$$

$= Y$ by definition.

Variance of Estimator

For convenience, let $\sigma_d^2 = \mathrm{Var}\ (\hat{Y}_d/E_\mu\ \delta(d,g))$ and σ_1 be the between-PSU variance for the original survey. Then

$$\mathrm{Var}\hat{Y} = \mathrm{Var}_\mu[E_{\lambda(g)}(\hat{Y}|g)] + \qquad (2)$$

$$E_\mu\{E_{\lambda(g)}[\mathrm{Var}(\hat{Y}|g,h)|g]\} + E_\mu\{\mathrm{Var}_{\lambda(g)}[E(\hat{Y}|g,h)|g]\}.$$

The first term of $(2)_2$ is the within-stratum-between-PSU variance, σ_1.

The second term of (2) is the within-PSU variance:

$$\text{term 2} = \sum_{d\in D} \sigma_d^2\ E_\mu\left(\frac{\delta(d,g)}{E_{\lambda(g)}\beta(d,h)}\right).$$

The third term of (2) is the between-strata variance:

$$\text{term 3} = E_\mu\left\{\mathrm{Var}_{\lambda(g)}\left[\sum_{d\in D}\frac{\delta(d,g)\beta(d,h)Y_d}{E_\mu\delta(d,g)E_{\lambda(g)}\beta(d,h)}\ |g\right]\right\}.$$

As we noted earlier, the within-stratum-between-PSU variance is fixed; no subsampling method can alter it. If Var $(\hat{Y})$ is to be minimized, then it must be by minimizing the within-PSU and between-strata variances. However, to minimize these terms, some information must be available on σ_d and Y_d, respectively, for all $d\in D$. While it is rare that information will be available on σ_d for all $d\in D$, it is common for information to be available on Y_d for all $d\in D$; that is, there is some ancillary characteristic X which is related in some manner to Y and for which X_d is known for all $d\in D$. (For example, Y could be current unemployment and X could be low-income housing as of the last census.) It is then clear that the within-PSU variance is uncontrollable and the only promising approach to minimizing $\mathrm{Var}(\hat{Y})$ is to minimize the between-strata variance of the ancillary characteristic:

$$E_\mu\left\{\mathrm{Var}_{\lambda(g)}\ [E\ (\hat{X}|g,h)|g]\right\}. \qquad (3)$$

Under the uninformed strategy, the chosen algorithm can be used to minimize

$E_\mu\left\{\mathrm{Var}_\nu[E(\hat{X}|g,h)|g]\right\}$ subject to the constraint (1). Let ν^* be this optimal ν. Because of the constraint, in most subsampling problems involving real as opposed to artificial populations, there will exist some ν' and $\tilde{g}$ such that

$$\mathrm{Var}_{\nu'}\left[E(\hat{X}|\tilde{g},h)|\tilde{g}\right] < \mathrm{Var}_{\nu^*}\left[E(\hat{X}|\tilde{g},h)|\tilde{g}\right].$$

It is then immediately clear that the informed strategy leads to a strictly better value of (3) than does the uninformed. Simply define

$$\lambda(g) = \begin{cases} \nu' \text{if } g = \tilde{g}, \\ \nu^* \text{ otherwise} \end{cases}$$

Even in the case where ν' and $\tilde{g}$ do not exist it is always true that the informed strategy leads to a value of (3) no worse than does the uninformed. In this case, just take $\lambda(g) = \nu$ for all $g\in G$.

Example

The example is given for an artificially simple situation to keep the calculations short. The purpose of the example is to reinforce the foregoing discussion. The reader should not take it as an illustration of the magnitude of the difference between the strategies. Suppose that there are 8 PSUs in D. The original survey formed four strata of two PSUs each. It furthermore used controlled selection[1] so that there are only two sets in G. For the new survey, we want to form two superstrata, select one stratum independently from each superstratum, with probability proportionate to a fixed stratum size, and then accept the sample PSUs in the selected strata. Table 1 provides the required parameters. The parameters Y_d and σ_d^2 are shown but assumed to be unknown. The known ancillary characteristic is X, the value of X for the d-th PSU is X_d.

In this situation, the uninformed strategy is to make the superstrata homogeneous with respect to

$$\sum_{d\in \text{stratum}} X_d/(\text{Stratum Size}).$$

The informed is to make the superstrata homogeneous with respect to

$$\sum_{d\in \text{stratum}} (\delta(d,g)X_d/E_\mu\delta(d,g))/(\text{Stratum Size}).$$

Table 2 gives these statistics for the strata. It is clear from this table that the best superstratification under the uninformed strategy is {1,2,}, {3,4,}.

Under the informed strategy, the best superstratification is {1,3}, {2,4} if $g=g_1$ and {1,2}, {3,4} if $g=g_2$. Table 3 shows the induced measures on H_g for each g.

We now calculate the variance on $\hat{Y}$. Table 4 contains some intermediate calculations for term 3 of (2).

Note that $\sum_{d\epsilon D} \frac{\delta(d,g_1)Y_d}{E_\mu \delta(d,g)} = 2561.43$ and $\sum_{d\epsilon D} \frac{\delta(d,g_2)Y_d}{E_\mu \delta(d,g)} = 4800.$

Thus (5) is (.7)(37177.2) + (.3)(6780.5) = 28058

Whereas (4) is (.7)(3057.5) + (.3)(6780.5) = 4174

For completeness, term 2 of (2) is either 325 for the uninformed strategy or 326 for the informed strategy. Term 1 of (2) is 1,236,393.

Thus the total variance of $\hat{Y}$ under the uninformed strategy is 1,264,776 versus 1,240,893 under the informed.

IV. VARIANCE ESTIMATION

We now focus on two specific sampling plans which are commonly used in practice and propose a reasonable variance estimator for each of the two sample designs. These are traditional variance estimators with modifications to suit each specific design. We have been unable, however, to give the precise expression for the expected values of the estimators under the informed strategy.

Sample Design I: The new survey forms superstrata of the strata and then selects one stratum per superstratum with probability proportionate to stratum size. The selection of strata is independent between superstrata.

Sample Design II: The same as sample design I except that 2 strata are selected with replacement from each superstratum.

As in the last section the within-PSU sampling is assumed to be independent from PSU to PSU and independent of the selection of sample PSUs. Without loss of generality, for variance estimation we will ignore the within-PSU sampling and consider only the cases where the true values are known at the PSU level.

It should be pointed out that the sample design I is being used by the Census Bureau for the redesigned sample of the American Housing Survey (AHS) which uses the informed strategy, while a sampling plan similar to the sample design II is being used for the General Purpose Sample (GPS) which uses the uninformed strategy. The Current Population Survey is the original survey for both these sample designs.

Before the derivation of variances and their estimators for these two specific sample designs, we need the following additional notation:

Let

L = total number of superstrata

K_i = total number of original-survey sample PSUs (or equivalently original-survey strata) in the i^{th} superstratum

$\hat{Y}_{ik}$ = the estimated total of characteristic Y based on the original-survey sample PSU for the k^{th} original survey stratum in the i^{th} superstratum (i.e., $\hat{Y}_{ik} = \frac{Y_d}{E_\mu \delta(d,g)}$ where d is the original-survey sample PSU.)

π_{ik} = probability of selecting the k^{th} original-survey sample PSU (or k^{th} original-survey stratum) within the i^{th} superstratum.

$\hat{Y}_i = \sum_{k=1}^{K_i} \hat{Y}_{ik}$, the i^{th} superstratum total estimated from all original-survey sample PSUs in the superstratum.

Sample Design I

Under this sample design, the estimator of the total of Y can be expressed as

$$\hat{Y}_I = \sum_{i=1}^{L} \frac{\hat{Y}_{id}}{\pi_{id}} \tag{4}$$

where d denotes the sample PSU selected by the new survey. Its variance, derived from equation (2) in Section III, is given by

$$\mathrm{Var}(\hat{Y}_I) = \sigma_1^2 + E_\mu \left\{ \sum_{i=1}^{L} \sum_{k=1}^{K_i} \pi_{ik} \left(\frac{\hat{Y}_{ik}}{\pi_{ik}} - \hat{Y}_i \right)^2 \right\} \tag{5}$$

where σ_1^2 is the between-PSU variance for the original survey as defined in Section III.

Since only one PSU was selected from a stratum in both phases of sampling, no unbiased estimator of $\mathrm{Var}(\hat{Y}_I)$ exists. The customary approach is to use collapsed superstrata to estimate variances.[2/]

Let observations in a typical h^{th} pair of superstrata be

$\frac{\hat{Y}_{h1}}{\pi_{h1}}, \frac{\hat{Y}_{h2}}{\pi_{h2}}$, where h goes from 1 to $\frac{L}{2}$. An estimator of Var $(\hat{Y}_I)$ can then be constructed as

$$\mathrm{var}(\hat{Y}_I) = \sum_{h=1}^{L/2} \left(\frac{\hat{Y}_{h1}}{\pi_{h1}} - \frac{\hat{Y}_{h2}}{\pi_{h2}} \right)^2 \tag{6}$$

Under the uninformed strategy, this estimator has a closed-form non-negative bias. Since the formation of superstrata in the informed strategy is dependent upon the outcome of PSU selection for the original survey, we have not been able to derive a satisfactory algebraic expression for the expected value of $\mathrm{var}(\hat{Y}_I)$ when this strategy is used. However, based on the form of this estimator we believe that $\mathrm{var}(\hat{Y}_I)$ in general overestimates $\mathrm{Var}(\hat{Y}_I)$. The bias may be reduced by pairing the superstrata based on superstrata totals of a correlated characteristics x_2. More definitive studies on the properties of $\mathrm{var}(\hat{Y}_I)$ under the informed strategy and the comparisons of these properties between the two strategies are needed.

Sample Design II

Let Z_{ik} be the selection probability of the k^{th} original-survey PSU in the i^{th} superstratum on each draw (i.e., sample size 1), then $\pi_{ik} = 2\,Z_{ik}$. Using the notation of this section, the unbiased estimator of the total of Y given in Section III can be written as

$$\hat{Y}_{II} = \sum_{i=1}^{L} \sum_{k=1}^{2} \frac{\hat{Y}_{ik}}{2Z_{ik}} \qquad (7)$$

According to equation (2) in Section III, the variance of $\hat{Y}_{II}$ can be written as

$$\mathrm{Var}(\hat{Y}_{II}) = \sigma_1^2 + E_\mu\left[\frac{1}{2} \sum_{i=1}^{L} \sum_{k=1}^{K_i} Z_{ik}\left(\frac{\hat{Y}_{ik}}{Z_{ik}} - \hat{Y}_i\right)^2\right] \quad (8)$$

It is easy to show that the second term in (8) can be unbiasedly estimated by

$$\frac{1}{4} \sum_{i=1}^{L}\left(\frac{\hat{Y}_{i1}}{Z_{i1}} - \frac{\hat{Y}_{i2}}{Z_{i2}}\right)^2$$

Since only one PSU is selected from each stratum for the original survey, it is not possible to obtain an unbiased estimator of σ_1^2 in (8). One cannot use the pair of original-survey strata within each superstratum because they are grouped into the same superstratum based on sample estimates. Such an approach would yield an underestimate of variance. However, one may pair strata from different superstrata based on superstrata totals of characteristics $\underset{\sim}{x}_2$. Let observations in a typical h^{th} pair be

$\frac{\hat{Y}_{h1k}}{Z_{h1k}}$ and $\frac{\hat{Y}_{h2k'}}{Z_{h2k'}}$ where h goes from 1 to L/2. $(k=1,2;\ k'=1,2)$

This leads to the following proposed variance estimator for $\mathrm{Var}(\hat{Y}_{II})$.

$$\mathrm{var}(\hat{Y}_{II}) = \frac{1}{2} \sum_{h=1}^{L/2}\left[\left(\frac{\hat{Y}_{h1k}}{Z_{h1k}} - \frac{\hat{Y}_{h2k'}}{Z_{h2k'}}\right)^2 + \left(\frac{\hat{Y}_{h1k'}}{Z_{h1k'}} - \frac{\hat{Y}_{h2k}}{Z_{h2k}}\right)^2\right] - \frac{1}{4} \sum_{i=1}^{L}\left(\frac{\hat{Y}_{i1}}{Z_{i1}} - \frac{\hat{Y}_{i2}}{Z_{i2}}\right)^2 \quad (9)$$

where superstrata h1 and h2 are paired together as described above. Within each superstrata pair, each of the two sample PSUs (or equivalently, original-survey strata) in one superstratum is randomly paired with one of the two sample PSUs in the other superstratum.

Under the uninformed strategy, it can be shown that the estimator has a closed-form non-negative bias. For the informed strategy, since the composition of superstrata are dependent upon the sample outcome of the original-survey selection and for the same reasons as stated for sample design I, we have not been able to show algebraically the bias of $\mathrm{var}(\hat{Y}_{II})$. However, we think that $\mathrm{var}(\hat{Y}_{II})$ will be a satisfactory estimator of $\mathrm{Var}(\hat{Y}_{II})$. Again, additional investigations on the properties of $\mathrm{var}(\hat{Y}_{II})$ under both strategies are needed.

V. CONCLUSION

This paper has compared two general strategies for stratification when it is desired to select a subset of sample PSUs from an original survey for a new survey. It was explained in Sections II and III why we expect that lower variance will result when sample PSU characteristics rather than stratum characteristics are used in forming strata and in other methods of subsampling. The easiest way to understand why this happens is to think in terms of double sampling and using the information from the first phase of sampling in the second phase. An example has also been given. Finally, variance estimators were provided for one sample PSU and two sample PSUs per superstratum in the new survey. For the case with two sample PSUs, an innovative approach was taken in which pairs of PSUs across different superstrata instead of from the same superstrata are used in the variance estimator in order to avoid underestimating the variance. We were unable, however, to derive expected values for the variance estimators.

This paper is of value in two ways. First, the results can be applied to use sample PSU characteristics rather than stratum characteristics for stratification and other methods of subsampling when a subset of sample PSUs is desired for a new survey. Second, it is instructive to understand why use of sample PSU characteristics is preferable for those readers whose intuition tells them otherwise.

ACKNOWLEDGMENTS
Beverley D. Causey at the Census Bureau made significant contributions to our early work in conceptualizing superstratification as a well-defined random event. We would also like to thank Joan George, Ovalecia High, and Edith Oechsler for doing the typing of this paper. Any errors in this paper are the sole responsibility of the authors.

---

1 Controlled selection is a procedure that goes beyond stratification in restricting the number of possible selection outcomes. Among other applications, it was used at the Bureau in the early 1970's to ensure representation of every state in the Current Population Survey. It is assumed here to keep the calculations short.

2 An apparently superior approach is given by Shapiro and Bateman (1979), but will not be discussed here.

TABLE 1.

d	Stratum	Stratum Size	X_d	Y_d	σ_d^2	$\delta(d,g_1)$	$\delta(d,g_2)$	$E_\mu\delta(d,g)$
1	1	1000	250	494	41	1	0	.7
2			250	515	32	0	1	.3
3	2	1200	240	493	51	1	0	.7
4			300	591	28	0	1	.3
5	3	800	200	429	36	1	0	.7
6			80	158	39	0	1	.3
7	4	900	180	377	42	1	0	.7
8			90	176	45	0	1	.3

TABLE 2.

Stratum	Homogeneity Measure		
	Uninformed Strategy	Informed Strategy	
		g_1	g_2
1	.500	.357	.833
2	.450	.286	.833
3	.350	.357	.333
4	.300	.286	.333

TABLE 3.

h (Set of Selected strata for survey 2)	$\nu(h)$	$[\lambda(g_1)](h)$	$[\lambda(g_2](h)$
{1,2}	0	.31746	0
{1,3}	.21390	0	.21390
{1,4}	.24064	.23810	.24064
{2,3}	.25668	.25397	.25668
{2,4}	.28877	0	.28877
{3,4}	0	.19048	0

TABLE 4.

	Uninformed Strategy		Informed Strategy	
	$E(\hat{Y}\mid g_1,h)$	$E(\hat{Y}\mid g_2,h)$	$E(\hat{Y}\mid g_1,h)$	$E(\hat{Y}\mid g_2,h)$
{1,2}	N/A	NA	2502.79	NA
{1,3}	2854.89	4895.83	NA	4895.83
{1,4}	2569.87	4884.81	2526.95	4884.81
{2,3}	2593.51	4730.83	2611.43	4730.83
{2,4}	2308.49	4719.81	NA	4719.81
{3,4}	N/A	NA	2635.60	NA