I. INTRODUCTION

There are situations where one wants to use a subset of the primary sample units (PSUs) in 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 high sampling and interviewing costs in PSUs not in the CPS sample PSUs. We might be interested in a subsample usually 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. (This information is used directly in the informed strategy when we make strata homogeneous with respect to the original survey population characteristics.) 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 sample PSU information we are given in the original survey 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.)

The issue to be discussed here is whether the use of sample PSU information for stratification is innovative and of potential application to other situations.
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 to take full advantage of the survey's stratification criteria. Let \( A \) be the set of all possible samples of size \( n \) from \( G \) and \( B \) be the set of all possible samples of size \( m \) from \( G \). Then the estimator of \( Y \) that we discuss is of the form:

\[
Y = \sum_{d \in D} \sum_{h \in H_d} \frac{n(h)}{m} \mu_d(h) Y_d(h) + \sum_{d \in D} \sum_{h \in H_d} \frac{n(h)}{m} \mu_d(h) Y_d(h)
\]

for some function \( \mu_d(h) \) that is defined on \( D \times H_d \) and depends only on the characteristics of the sample PSU. The function \( \mu_d(h) \) may be interpreted as the probability of selecting a sample PSU in \( h \) given that it was selected for the original survey. 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, \( Y_d(h) \) is the probability that the strata represented by the sample PSUs \( h \) 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 \( h \). Given this constraint on \( Y_d(h) \), we simply write \( Y \) for \( Y_d(h) \).

To summarize the notation developed so far:

\[
\begin{align*}
D &= \text{All PSUs}, \\
G &= \text{All possible sets of original-survey sample PSUs}, \\
\nu &= \text{Original survey measure on } G, \\
H_d &= \text{All possible sets of new-survey sample PSUs given } g \in G \\
\lambda(g) &= \text{Informed strategy measure on } H_d \\
\nu &= \text{uninformed strategy measure on } H_d \\
\end{align*}
\]

**Estimator Definition**

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

\[
\begin{align*}
\psi(d,g) &= \begin{cases} 1 \text{ if } d \cap g \neq \emptyset, \\
0 \text{ otherwise,} \\
\end{cases} \\
\Phi(d,h) &= \begin{cases} 1 \text{ if } d \cap h \neq \emptyset, \\
0 \text{ otherwise.} \\
\end{cases}
\end{align*}
\]

Then the estimator of \( Y \) that we discuss is
\[ \bar{Y} = \frac{\sum_{d \in D} \bar{Y}_d \mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\sum_{d \in D} \mu_{d}(g) \mu_{d}(h)} \] for the selected \( g \in G \) and the selected \( h \in G \).

Note that \( \mu_{d}(g) \) is the probability of selecting PSU \( d \) for the original survey and that \( \mu_{d}(g) \) 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 \( \mu_{d}(g) > 0 \) \( \forall d \in D \) and \( \mu_{d}(g) \mu_{d}(h) > 0 \) \( \forall d \in D \) and \( \forall g \in G \) such that \( \mu_{d}(g) \mu_{d}(h) = 1 \).

**Proof of Unbiasedness of Estimator**

Since the within-PSU sampling is independent of the PSU sampling and \( \mu_{d}(g) \) is fixed given \( g \), we have that

\[ E\{Y\} = \sum_{d \in D} Y_d \mu_{d}(g) \mu_{d}(h) \mu_{d}(g) \mu_{d}(h) \mu_{d}(g) \mu_{d}(h) \]

\[ = \sum_{d \in D} Y_d \mu_{d}(g) \mu_{d}(h) \mu_{d}(g) \mu_{d}(h) \mu_{d}(g) \mu_{d}(h) \]

\[ = Y \] by definition.

**Variance of Estimator**

For convenience, let \( \sigma^2 = \text{Var}\{Y\}/E\{Y_d\} \) and \( \sigma^2 \) be the between-PSU variance for the original survey. Then

\[ \text{Var}\{\bar{Y}\} = \text{Var}\{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d\} + \text{Var}\{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d\} \]

The first term of (2) is the within-stratum-between-PSU variance, \( \sigma^2 \).

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

\[ \text{term 2} = \sum_{d \in D} \frac{\sigma^2}{\mu_{d}(g) \mu_{d}(h)} \]

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

\[ \text{term 3} = \sum_{d \in D} \frac{\sigma^2}{\mu_{d}(g) \mu_{d}(h)} \sum_{d \in D} \frac{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d} \]

As we noted earlier, the within-stratum-between-PSU variance is fixed; no subsampling method can alter it. If Var\{\bar{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 \( \sigma^2 \) and \( \sigma^2 \), respectively, for all \( d \in D \). While it is rare that information will be available on \( \sigma^2 \) for all \( d \in D \), it is common for information to be available on \( \bar{Y}_d \) for all \( d \in D \); that is, there is some ancillary characteristic \( X \) which is related in some manner to \( Y \) and for which \( X \) 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 Var\{\bar{Y}\} is to minimize the between-strata variance of the ancillary characteristic:

\[ \text{Var}_{\lambda}(g) \left( \sum_{d \in D} \frac{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d} \right) \]

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

\[ \text{Var}_{\lambda}(g) \left( \sum_{d \in D} \frac{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d} \right) \]

subject to the constraint

\[ \lambda(g) \{ \nu \} \]

Let \( \nu^* \) be this optimal \( \nu \). Because of the constraint, in most subsampling problems involving real as opposed to artificial populations, there will exist some \( \nu^* \) and \( g \) such that

\[ \text{Var}_{\lambda}(g) \left( \sum_{d \in D} \frac{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d} \right) < \text{Var}_{\lambda}(g) \left( \sum_{d \in D} \frac{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d} \right) \]

Even in the case where \( \nu^* \) and \( g \) do not exist it is always true that the informed strategy leads to a strictly better value of (3) than does the uninformed. Simply define

\[ \lambda(g) = \{ \nu^* \} \]

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

\[ \text{Var}_{\lambda}(g) \left( \sum_{d \in D} \frac{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d}{\mu_{d}(g) \mu_{d}(h) \bar{Y}_d} \right) \]

subject to the constraint

\[ \lambda(g) \{ \nu \} \]

Example

The example is given for an artificially simple situation to keep the calculations short. The purpose of the example is to reinforce the preceding 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 PSU each. It furthermore used controlled selection 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 \( \bar{Y}_d \) and \( \mu_{d}(g) \) 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

\[ \sum_{d \in D} X_d \text{(Stratum Size)} \]

The informed is to make the superstrata homogeneous with respect to

\[ \sum_{d \in D} X_d \text{(Stratum Size)} \]

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 \( g = g \), and \( \{1,2\}, \{3,4\} \) if \( g = g \). Table 3 shows the induced measures on \( H \) for each \( g \).

We now calculate the variance on \( \bar{Y} \). Table 4 contains some intermediate calculations for term 3 of (2).
Note that \( \sum_{d \in D} \frac{\epsilon(d,g_1)Y_d}{\mu} = 2561.43 \) and
\( \sum_{d \in D} \frac{\epsilon(d,g_2)Y_d}{\mu} = 4800. \)
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 \( \hat{\gamma} \) 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.

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
\( \hat{\gamma}_{ik} = \) 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., \( \hat{\gamma}_{ik} = \frac{\sum_{d \in D} \epsilon(d,g)Y_d}{\mu} \) where \( d \) is the original-survey sample PSU.)

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

\( \hat{\gamma}_I = \sum_{i=1}^{L} \sum_{k=1}^{K_i} \hat{\gamma}_{ik} \), the \( i \)-th superstratum total estimated from all original-survey sample PSUs in the superstratum.

#### Sample Design II

Under this sample design, the estimator of the total of \( Y \) can be expressed as
\[ \hat{\gamma}_I = \frac{1}{2} \sum_{i=1}^{L} \hat{\gamma}_{id} \] (4)
where \( d \) denotes the sample PSU selected by the new survey. Its variance, derived from equation (2) in Section III, is given by
\[ \text{Var}(\hat{\gamma}_I) = \sigma^2 + \sum_{h=1}^{L} \sum_{k=1}^{K_h} \left( \frac{\hat{\gamma}_{h1}}{\pi_{h1}} - \frac{\hat{\gamma}_{h2}}{\pi_{h2}} \right)^2 \] (5)

where \( \sigma^2 \) is the between-PSU variance for the original survey as defined in Section III.

Since only one PSU was selected from each stratum in both phases of sampling, no unbiased estimator of \( \text{Var}(\hat{\gamma}_I) \) exists. The customary approach is to use collapsed superstrata to estimate variances. Let observations in a typical \( h \)-th pair of superstrata be
\[ \frac{\hat{\gamma}_{h1}}{\pi_{h1}}, \frac{\hat{\gamma}_{h2}}{\pi_{h2}} \]
where \( h \) goes from 1 to \( \frac{L}{2} \). An estimator of \( \text{Var}(\hat{\gamma}_I) \) can then be constructed as
\[ \text{Var}(\hat{\gamma}_I) = \sum_{h=1}^{L/2} \left( \frac{\hat{\gamma}_{h1}}{\pi_{h1}} - \frac{\hat{\gamma}_{h2}}{\pi_{h2}} \right)^2 \] (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 \( \text{Var}(\hat{\gamma}_I) \) when this strategy is used. However, based on the form of this estimator we believe that \( \text{Var}(\hat{\gamma}_I) \) in general overestimates \( \text{Var}(\hat{\gamma}_I) \). The bias may be reduced by pairing the superstrata based on superstrata totals of a correlated characteristic \( x \). More definitive studies on the properties of \( \text{Var}(\hat{\gamma}_I) \) under the informed strategy and the comparisons of these properties between the two strategies are needed.

#### Sample Design II

Let \( \pi_{ik} \) 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 \( \pi_{ik} = 2 \pi_{ik} \). Using the notation of this section, the unbiased estimator of the total of \( Y \) given in Section III can be written as
According to equation (2) in Section III, the variance of $\bar{Y}_{II}$ can be written as

$$\text{Var}(\bar{Y}_{II}) = \sigma^2 + \frac{1}{L} \sum_{i=1}^{L} \sum_{k=1}^{K} \left( \frac{\bar{Y}_{ik} - \bar{Y}_{II}}{Z_{ik}} \right)^2$$

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

$$\frac{1}{L} \sum_{i=1}^{L} \left( \frac{\bar{Y}_{11} - \bar{Y}_{12}}{Z_{11}} \right)^2,$$

Since only one PSU is selected from each stratum for the original survey, it is not possible to obtain an unbiased estimator of $\sigma^2$ 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 $x_2$. Let observations in a typical $h$th pair be $\bar{Y}_{hlk}$ and $\bar{Y}_{h2k}$, where $h$ goes from 1 to $L/2$. This leads to the following proposed variance estimator for $\text{Var}(\bar{Y}_{II})$.

$$\text{var}(\bar{Y}_{II}) = \frac{1}{L} \sum_{h=1}^{L/2} \left( \frac{\bar{Y}_{hlk} - \bar{Y}_{h2k}}{Z_{hlk} - Z_{h2k}} \right)^2$$

where superstrata $h_1$ and $h_2$ 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 $\text{var}(\bar{Y}_{II})$. However, we think that $\text{var}(\bar{Y}_{II})$ will be a satisfactory estimator of $\text{Var}(\bar{Y}_{II})$. Again, additional investigations on the properties of $\text{var}(\bar{Y}_{II})$ 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.
### Table 1.

<table>
<thead>
<tr>
<th>d</th>
<th>Stratum</th>
<th>Stratum Size</th>
<th>$X_d$</th>
<th>$Y_d$</th>
<th>$\sigma_d$</th>
<th>$\sigma(d,g_1)$</th>
<th>$\sigma(d,g_2)$</th>
<th>$E(\sigma(d,g))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1000</td>
<td>250</td>
<td>494</td>
<td>41</td>
<td>1</td>
<td>0</td>
<td>.7</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1200</td>
<td>250</td>
<td>515</td>
<td>32</td>
<td>0</td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>800</td>
<td>300</td>
<td>591</td>
<td>51</td>
<td>0</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>900</td>
<td>300</td>
<td>591</td>
<td>28</td>
<td>0</td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>300</td>
<td>300</td>
<td>591</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>.7</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>200</td>
<td>200</td>
<td>429</td>
<td>45</td>
<td>0</td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>180</td>
<td>180</td>
<td>377</td>
<td>42</td>
<td>1</td>
<td>0</td>
<td>.7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>90</td>
<td>90</td>
<td>176</td>
<td>45</td>
<td>0</td>
<td>1</td>
<td>.3</td>
</tr>
</tbody>
</table>

### Table 2.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Homogeneity Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uninformed Strategy</td>
</tr>
<tr>
<td></td>
<td>$\delta_1$</td>
</tr>
<tr>
<td>1</td>
<td>.500</td>
</tr>
<tr>
<td>2</td>
<td>.450</td>
</tr>
<tr>
<td>3</td>
<td>.350</td>
</tr>
<tr>
<td>4</td>
<td>.300</td>
</tr>
</tbody>
</table>

### Table 3.

<table>
<thead>
<tr>
<th>$h$ (Set of Selected strata for survey 2)</th>
<th>$v(h)$</th>
<th>$<a href="h">\lambda(g_1)</a>$</th>
<th>$<a href="h">\lambda(g_2)</a>$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1, 2}</td>
<td>0</td>
<td>.31766</td>
<td>0</td>
</tr>
<tr>
<td>{1, 3}</td>
<td>.21390</td>
<td>0</td>
<td>.21390</td>
</tr>
<tr>
<td>{1, 4}</td>
<td>.24064</td>
<td>.23810</td>
<td>.24064</td>
</tr>
<tr>
<td>{2, 3}</td>
<td>.25668</td>
<td>.25397</td>
<td>.25668</td>
</tr>
<tr>
<td>{2, 4}</td>
<td>.28877</td>
<td>0</td>
<td>.28877</td>
</tr>
<tr>
<td>{3, 4}</td>
<td>0</td>
<td>.19048</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 4.

| $h$ (Set of Selected strata for survey 2) | $E(\tilde{Y}|g_1,h)$ | $E(\tilde{Y}|g_2,h)$ | $E(\tilde{Y}|g_1,h)$ | $E(\tilde{Y}|g_2,h)$ |
|---------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| {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                    |