Conditional Properties of Post-Stratified

Estimators Under Normal Theory

Robert J. Casady
Richard Valliant

U.S. Bureau of Labor Statistics
441 G St. NW
Washington DC 20212

Revised 10/26/92

This paper was prepared for presentation at the 1992 convention of the American Statistical Association held in Boston MA. Any opinions expressed are those of the authors and do not reflect policy of the Bureau of Labor Statistics.

CONDITIONAL PROPERTIES OF POST-STRATIFIED ESTIMATORS UNDER NORMAL THEORY

ROBERT J. CASADY and RICHARD VALLIANT[1]

ABSTRACT

Post-stratification is a common technique for improving precision of estimators by using data items not available at the design stage of a survey. In large, complex samples, the vector of Horvitz-Thompson estimators of survey target variables and of post-stratum population sizes will, under appropriate conditions, be approximately multivariate normal. This large sample normality leads to a new post-stratified regression estimator, which is analogous to the linear regression estimator in simple random sampling. We derive the large sample design bias and mean squared errors of this new estimator, the standard post-stratified estimator, the Horvitz-Thompson estimator, and a ratio estimator. We use both real and artificial populations to study empirically the conditional and unconditional properties of the estimators in multistage sampling.

KEYWORDS: Asymptotic normality; Regression estimator; Defective frames; Ratio estimator; Horvitz-Thompson estimator.

1. INTRODUCTION

1.1 Background

A major thrust in sampling theory in the last twenty years has been to devise ways of restricting the set of samples used for inference. In a purely design-based approach, as described in Hansen, Madow, and Tepping (1983), no such restrictions are imposed. Statistical properties are calculated by averaging over the set of all samples that might have been selected using a particular design. Although it is generally conceded that some type of design-based, conditional inference is desirable (Fuller 1981, Rao 1985, Hidiroglou and Särndal 1989), satisfactory theory has yet to be developed except in relatively simple cases. Alternative approaches are prediction theory, developed by Royall (1971) and many others, and the Bayesian approach, found in Ericson (1969), which avoid averaging over repeated samples through the use of superpopulation models.

[1] Robert J. Casady and Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachuetts Ave. N.E., Washington D.C., 20212-0001.

A design-based approach to conditioning was introduced by Robinson (1987) for the particular case of ratio estimates in sample surveys. Robinson applied large sample theory and approximate normality of certain statistics to produce a conditional, design-based theory for the ratio estimator.

In this paper, we extend that line of reasoning to the problem of post-stratification. Convincing arguments have been made in the past by Durbin (1969), Holt and Smith (1979), and Yates (1960) that post-stratified samples should be analyzed conditional on the sample distribution of units among the post-strata. However, as Rao (1985) has noted, the difficulties in developing an exact, design-based, finite sample theory for post-stratification in general sample designs may be intractable. Model-based, conditional analyses of post-stratified samples are presented in Little (1991) and Valliant (1993). The alternative pursued here is design-based and uses large sample, approximate normality in a way similar to that of Robinson (1987) as a means studying conditional properties of estimators.

1.2 Basic Definitions and Notation

The **target population** is a well defined collection of elementary (or analytic) units. For many applications the elementary units are either persons or establishments. We assume the target population has been partitioned into **first stage sampling units** (FSUs). For person based surveys the FSUs are commonly households, groups of households or even counties, while for establishment based surveys it is not uncommon that the individual establishment is an FSU. In any event, the collection of FSUs will be referred to as the **first stage sampling frame** (or just **sampling frame**). It is assumed that there are M FSUs in the sampling frame and they are labeled 1, 2, ..., M. We also assume that the population units can be partitioned into K "post-strata" which can be used for the purposes of estimation.

We let y represent the value of the characteristic of interest (e.g. weekly income, number of hours worked last week, restricted activity days in last two weeks, etc.) for an elementary unit. Associated with the i^{th} FSU are $2K$ real numbers:

y_{ik} = aggregate of the y values for the elementary units in the i^{th} FSU which are in the k^{th} post-stratum,

N_{ik} = number of elementary units in the i^{th} FSU which are in the k^{th} post-stratum.

For each post-stratum we then define

$Y_{\cdot k} = \sum_{i=1}^{M} y_{ik}$ = aggregate of the y values for all elementary units in the k^{th} post-stratum,

$N_{\cdot k} = \sum_{i=1}^{M} N_{ik}$ = total number of elementary units in the k^{th} post-stratum.

In what follows we assume that the $N_{\cdot k}$ are known. The population aggregate of the y values is given by $Y_{\cdot\cdot} = \sum_{k=1}^{K} Y_{\cdot k}$ and the total population size by $N_{\cdot\cdot} = \sum_{k=1}^{K} N_{\cdot k}$. In sections 1-3, we assume that the sampling frame provides "coverage" of the entire target population. In section 4, we consider the problem of a defective frame, i.e. one in which the coverage of the frame differs from that of the target population.

1.3 Sample Design and Basic Estimation

Suppose that the first stage sampling frame is partitioned into L strata and that a multi-stage, stratified design is used with a total sample of m FSUs. In the following, the subscript representing design strata is suppressed in order to simplify the notation. For the subsequent theory, it is unnecessary to explicitly define sampling and estimation procedures for second and higher levels of the design. However, for every sample FSU, we require estimators $\hat{y}_{ik}$ and $\hat{N}_{ik}$ so that $\underset{2+}{\mathrm{E}}\left[\hat{y}_{ik}\right] = y_{ik}$ and $\underset{2+}{\mathrm{E}}\left[\hat{N}_{ik}\right] = N_{ik}$ where the notation $\underset{2+}{\mathrm{E}}$ indicates the design-expectation over stages 2 and higher. Letting π_i be the probability that the i^{th} FSU is included in the sample and $w_i = 1/\pi_i$, it follows that the estimator $\hat{Y}_k = \sum_{i=1}^{m} w_i \hat{y}_{ik}$ is unbiased for Y_k and the estimator $\hat{N}_k = \sum_{i=1}^{m} w_i \hat{N}_{ik}$ is unbiased for N_k.

1.4 An Analogue to Robinson's Asymptotic Result

Following Krewski and Rao (1981), we can establish our asymptotic results as $L \to \infty$ within in the framework of a sequence of finite populations $\{\Pi_L\}$ with L strata in Π_L. It should be understood that we implicitly assume (without formal statement) the sample design and regularity conditions as specified in Krewski and Rao and more fully developed in Rao and Wu (1985). Details of proofs add little to those in the literature and are omitted.

Converting to matrix notation, we let $\mathbf{Y} = [Y_1 \cdots Y_{.K}]'$, $\mathbf{N} = [N_1 \cdots N_{.K}]'$, $\hat{\mathbf{Y}} = [\hat{Y}_1 \cdots \hat{Y}_{.K}]'$, $\hat{\mathbf{N}} = [\hat{N}_1 \cdots \hat{N}_K]'$ and $\mathbf{V} = \mathrm{var}\left\{\left[\bar{\hat{\mathbf{Y}}} \quad \bar{\hat{\mathbf{N}}}\right]'\right\}$ where $\bar{\hat{\mathbf{Y}}} = (1/N)\hat{\mathbf{Y}}$ and $\bar{\hat{\mathbf{N}}} = (1/N)\hat{\mathbf{N}}$. Note that $\bar{\hat{\mathbf{Y}}}$, which uses N in the denominator, is a notational convenience and does estimate means in the post-strata. Analogous to conditions C4 and C5 of Krewski and Rao (1981), we assume that

$$\lim_{L \to \infty} \frac{Y_k}{N_k} = \mu_k \text{ , for } k = 1, 2, \cdots, K \text{ ,} \tag{1}$$

$$\lim_{L \to \infty} \frac{N_k}{N} = \phi_k > 0 \text{ for } k = 1, 2, \cdots, K \text{ , and} \tag{2}$$

$$\lim_{L \to \infty} m\mathbf{V} = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{ (positive definite)} \tag{3}$$

where Σ is partitioned in the obvious manner. Note that we have again suppressed the subscript representing design strata. Assumptions (1)-(3) simply require that certain key quantities stabilize in large populations. Condition (2), in particular, assures that no post-stratum is empty as the population size increases. We now state the following,

Result: Assume the sample design and regularity conditions specified in Krewski and Rao and that Σ_{22}^{-1} exists; then, given $\bar{\hat{\mathbf{N}}}$, the conditional distribution of $\bar{\hat{\mathbf{Y}}}$ is asymptotically $N\left(\mathbf{M}_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\hat{\bar{\mathbf{N}}} - \mathbf{M}_2\right), m^{-1}\mathbf{V}_c\right)$, where $\mathbf{V}_c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $\mathbf{M}_1 = \lim_{L\to\infty}\hat{\bar{\mathbf{Y}}} = \left[\phi_1\mu_1 \quad \phi_2\mu_2 \quad \cdots \quad \phi_K\mu_K\right]'$ and $\mathbf{M}_2 = \lim_{L\to\infty}\hat{\bar{\mathbf{N}}} = \left[\phi_1 \quad \phi_2 \quad \cdots \quad \phi_K\right]'$.

Proof. This result is analogous to the result for K=1 given by Robinson (1987) and follows directly from the fact that the random vector $m^{1/2}\begin{bmatrix}\hat{\bar{\mathbf{Y}}} - \mathbf{M}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\hat{\bar{\mathbf{N}}} - \mathbf{M}_2) \\ \hat{\bar{\mathbf{N}}} - \mathbf{M}_2\end{bmatrix}$ tends in distribution to $N\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{V}_c & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}\end{bmatrix}\right)$. Strictly, as in Robinson, we consider the conditional distribution of $\hat{\bar{\mathbf{Y}}}$ for $\hat{\bar{\mathbf{N}}}$ in a cell of size $\varepsilon m^{-1/2}$ for small ε. . Note that in some sample designs $\mathbf{1}'\hat{\mathbf{N}} = N$ (such as those in which a fixed number of elementary units are selected with equal probabilities) in which case Σ_{22}^{-1} does not exist; in such cases only the first K-1 post-strata are considered for the purpose of conditioning.

In the next section, the asymptotic mean of $\hat{\bar{\mathbf{Y}}}$ is used to motivate a linear regression estimator of the population mean of the *y*'s.

2. CONDITIONAL PROPERTIES OF ESTIMATORS FOR THE POPULATION MEAN

2.1 Estimators for the Population Mean

The **population mean** is, by definition, $\mu = \lim_{L \to \infty}(Y/N)$ = $\lim_{L \to \infty}(\mathbf{1}'\mathbf{Y}/\mathbf{1}'\mathbf{N}) = \sum_{k=1}^{K} \phi_k \mu_k$ where $\mathbf{1}'$ is a row vector of K ones. Four estimators of the population mean will be considered. The first three are standard estimators found in the literature while the fourth is a new estimator motivated by the asymptotic, joint normality of $\hat{\bar{\mathbf{Y}}}$ and $\hat{\bar{\mathbf{N}}}$:

(1) Horvitz-Thompson estimator

$$\hat{\bar{Y}}_{HT} = \mathbf{1}'\hat{\mathbf{Y}} \big/ \mathbf{1}'\mathbf{N} = \mathbf{1}'\hat{\bar{\mathbf{Y}}}$$

(2) ratio estimator

$$\hat{\bar{Y}}_{R} = \mathbf{1}'\hat{\mathbf{Y}} \big/ \mathbf{1}'\hat{\mathbf{N}} = \mathbf{1}'\hat{\bar{\mathbf{Y}}} \big/ \mathbf{1}'\hat{\bar{\mathbf{N}}}$$

(3) post-stratified estimator

$$\hat{\bar{Y}}_{PS} = N^{-1} \sum_{k=1}^{K} \left(\frac{N_k}{\hat{N}_{.k}} \right) \hat{Y}_k = \mathbf{r}'\hat{\bar{\mathbf{Y}}} \text{ where } \mathbf{r}' = \left[N_1 \big/ \hat{N}_1, \ldots, N_K \big/ \hat{N}_K \right]$$

(4) linear regression estimator

$$\hat{\bar{Y}}_{LR} = \left[\mathbf{1}' \left(\hat{\bar{\mathbf{Y}}} - \Sigma_{12}\Sigma_{22}^{-1} \left(\hat{\bar{\mathbf{N}}} - \mathbf{M}_2 \right) \right) \right]$$

The linear regression estimator is motivated by the form of the large sample mean of the conditional random variable $\hat{\bar{\mathbf{Y}}} \big| \hat{\bar{\mathbf{N}}}$ listed at the end of section 1.4 and is very similar to the generalized regression estimator discussed by Särndal, Swensson and Wretman (1992). The linear regression estimator (4) was also discussed in the context of calibration estimation by Rao (1992). It should be noted that the ratio estimator does not require that N_k or their sum N be known. The Horvitz-Thompson estimator only requires that N be known, whereas the post-stratified and linear regression estimators require that $\{N_k | k = 1, \ldots, K\}$ be known. In practice, the linear regression estimator has the additional complication that the covariance matrices Σ_{12} and Σ_{22} are unknown and must be estimated from the sample. In implementing $\hat{Y}_{LR}$ in section 3, the known finite population quantities $(1/N)\mathbf{N}$ will be used in place of the limiting vector $\mathbf{M}_2$.

2.2 Conditional expectations and variances of the estimators

Using the asymptotic setup given earlier, the conditional expectations and variances of the four estimators can be computed. First, define the following three matrices:

$\mathbf{H} = \Sigma_{12}\Sigma_{22}^{-1}$,

$\mathbf{R} = \mathbf{H} - \mathbf{D}(\mu)$, and

$\mathbf{P} = \mathbf{H} - \mathbf{D}(\mu_k)$

where $\mathbf{D}(\mu) = diag(\mu, \ldots, \mu)$ and $\mathbf{D}(\mu_k) = diag(\mu_1, \ldots, \mu_K)$ are $K \times K$ diagonal matrices . Below, we state the mean and variance of the four estimators without providing any details of the calculations. When the sample of first-stage units is large, each of the estimators has essentially the same conditional variance. The Horvitz-Thompson, ratio, and post-stratified estimators are, however, conditionally biased, whereas the linear regression estimator is not. Thus, the linear regression estimator has the smallest asymptotic mean square error among the four estimators considered here. Rao (1992) also noted the optimality of the regression estimator within a certain class of difference estimators and its negligible large sample bias.

(1) Horvitz-Thompson estimator:

$$\mathrm{E}\left[\hat{Y}_{HT}\middle|\hat{\mathbf{N}}\right] = \mu + \left[\mathbf{1}'\mathbf{H}\left(\hat{\mathbf{N}} - \mathbf{M}_2\right)\right]$$

$$\mathrm{var}\left[\hat{Y}_{HT}\middle|\hat{\mathbf{N}}\right] = m^{-1}\left[\mathbf{1}'\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)\mathbf{1}\right] = m^{-1}\left[\mathbf{1}'\mathbf{V}_{\mathbf{c}}\mathbf{1}\right] = V_{HT(\mathbf{c})}$$

(2) ratio estimator:

$$\mathrm{E}\left[\hat{Y}_{R}\middle|\hat{\mathbf{N}}\right] = \mu + \left(\frac{N}{\hat{N}}\right)\left[\mathbf{1}'\mathbf{R}\left(\hat{\mathbf{N}} - \mathbf{M}_2\right)\right]$$

$$= \mu + \left[\mathbf{1}'\mathbf{R}\left(\hat{\mathbf{N}} - \mathbf{M}_2\right)\right] + o\left(m^{-1}\right)$$

$$\operatorname{var}\left[\hat{\bar{Y}}_R \middle| \hat{\bar{\mathbf{N}}}\right] = \left(N / \hat{N}\right)^2 V_{HT(\mathbf{c})}$$

$$= V_{HT(\mathbf{c})} + o\left(m^{-(3/2)}\right)$$

(3) post-stratified estimator:

$$\mathrm{E}\left[\hat{\bar{Y}}_{PS} \middle| \hat{\bar{\mathbf{N}}}\right] = \mu + \left[\mathbf{r}'\mathbf{P}\left(\hat{\bar{\mathbf{N}}} - \mathbf{M}_2\right)\right]$$

$$= \mu + \left[\mathbf{1}'\mathbf{P}\left(\hat{\bar{\mathbf{N}}} - \mathbf{M}_2\right)\right] + o\left(m^{-1}\right)$$

$$\operatorname{var}\left[\hat{\bar{Y}}_{PS} \middle| \hat{\bar{\mathbf{N}}}\right] = m^{-1}\left[\mathbf{r}'\mathbf{V}_{\mathbf{c}}\mathbf{r}\right]$$

$$= V_{HT(\mathbf{c})} + o\left(m^{-(3/2)}\right)$$

(4) linear regression estimator:

$$\mathrm{E}\left[\hat{\bar{Y}}_{LR} \middle| \hat{\bar{\mathbf{N}}}\right] = \mu$$

$$\operatorname{var}\left[\hat{\bar{Y}}_{LR} \middle| \hat{\bar{\mathbf{N}}}\right] = V_{HT(\mathbf{c})}$$

As noted in section 1, some minor modifications of the above formulas are necessary for designs, such as simple random sampling, in which $\mathbf{1}'\hat{\bar{\mathbf{N}}} = N$. The derivation of the requisite modifications is straightforward and is not detailed here.

The large-sample biases of the first three estimators depend on $\hat{\bar{\mathbf{N}}} - \mathbf{M}_2$. In other words, their biases are determined by how well the sample estimates the population distribution among the post-strata. In some special cases each of the first three can be conditionally unbiased. The post-stratified estimator, for example, will be approximately unbiased if $\mathbf{1}'\left(\mathbf{H} - \mathbf{D}(\mu_k)\right) = \mathbf{0}'$. This occurs in simple random sampling and is possible, though certainly not generally true, in more complex designs. The matrix $\mathbf{H}$ can be interpreted as the slope in a multivariate regression of $\hat{\bar{\mathbf{Y}}}$ on $\hat{\bar{\mathbf{N}}}$, or of $\bar{\mathbf{Y}}$ on $\bar{\mathbf{N}}$ when the sample estimates are close to the population values. Thinking heuristically in superpopulation terms, if $E_\xi\left(y_{ik}\right) = \mu_k N_{ik}$, as in Valliant (1993), with E_ξ denoting an expectation with respect to the model, then $E_\xi\left(Y_k\right) = \mu_k N_k$. The slope of the regression of Y_k on N_k is then μ_k and, in the unusual case in which the $\hat{Y}_k$'s are independent, $\mathbf{H}$ is diagonal. In fact $\mathbf{H} = \mathbf{D}(\mu_k)$, so the conditional design-bias of the post-stratified estimator would be zero. If, on the other hand, the model has an intercept, i.e. if $E_\xi(Y_k) = \alpha_k + \mu_k N_k$, then the post-stratified estimator may have a substantial conditional design-bias. We will use this line of reasoning in the empirical study in section 3 to devise a population for which $\hat{Y}_{ps}$ is conditionally biased.

Similar model-based thinking can be applied to the Horvitz-Thompson and ratio estimators to identify populations where the conditional design-biases will be predictably small for large samples. Suppose, as above, that the $\hat{Y}_k$'s are independent. If each post-stratum total is unrelated to the number of units in the post-stratum, i.e. a peculiar situation in which $E_\xi(Y_k)$ does not depend on N_k, then $\hat{Y}_{HT}$ is conditionally design-unbiased. If $E_\xi(Y_k) = \mu N_k$, implying that all elementary population units have the same mean regardless of post-stratum, then $\hat{Y}_R$ is conditionally design-unbiased.

2.3 Unconditional expectations and variances of the estimators

Unconditionally, all estimators are approximately design-unbiased as noted below. The relative sizes of the variances depend on the values of Σ_{12}, Σ_{22}, μ, and $\mathbf{D}(\mu_k)$. This is similar to the case of simple random sampling of a target *y* and an auxiliary *x*. In that case, whether the ratio estimator, $\bar{y}_s\,\bar{x}/\bar{x}_s$, or the regression estimator, $\bar{y}_s + b(\bar{x} - \bar{x}_s)$, has smaller design-variance also depends on the values of certain population parameters.

(1) Horvitz-Thompson estimator:

$$\mathrm{E}\left[\hat{Y}_{HT}\right] = \mu$$

$$\mathrm{var}\left[\hat{Y}_{HT}\right] = m^{-1}\left[\mathbf{1}'\Sigma_{11}\mathbf{1}\right]$$

(2) ratio estimator:

$$\mathrm{E}\left[\hat{Y}_R\right] = \mu + o\left(m^{-1}\right)$$

$$\mathrm{var}\left[\hat{Y}_R\right] = m^{-1}\left[\mathbf{1}'\left[\Sigma_{11} - 2\mu\Sigma_{21} + \mu^2\Sigma_{22}\right]\mathbf{1}\right] + o\left(m^{-(3/2)}\right)$$

(3) post-stratified estimator:

$$\mathrm{E}\left[\hat{Y}_{PS}\right] = \mu + o\left(m^{-1}\right)$$

$$\mathrm{var}\left[\hat{Y}_{PS}\right] = m^{-1}\left[\mathbf{1}'\left[\Sigma_{11} - 2\mathbf{D}\left(\mu_k\right)\Sigma_{21} + \mathbf{D}\left(\mu_k\right)\Sigma_{22}\mathbf{D}\left(\mu_k\right)\right]\mathbf{1}\right] + o\left(m^{-(3/2)}\right)$$

(4) linear regression estimator:

The unconditional expectation and variance are the same as the conditional expectation and variance.

3. SIMULATION RESULTS

The theory developed in the preceding sections was tested in a set of simulation studies using three separate populations. The population size and basic sample design parameters for the three studies are listed in Table 1. The first population consists of a subset of the persons included in the first quarter sample of the 1985 National Health Interview Survey (NHIS) and the second population consists of a subset of the persons included in the September 1988 sample from the Current Population Survey (CPS). Both the NHIS and CPS are sample surveys conducted by the U.S. government. The variable of interest for the NHIS population is the number of restricted activity days in the two weeks prior to the interview and the variable of interest for the CPS population is weekly wages per person.

Post-strata in the NHIS and CPS populations were formed on the basis of demographic characteristics (as is typically done in household surveys) in order to create population sub-groups that were homogenous with respect to the variable of interest. For the NHIS population the variables age and sex were used to define 4 post-strata and for the CPS population the variables age, race, and sex were used to define 8 post-strata.

The third population is artificial; it was created with the intention of producing a substantial conditional bias in the post-stratified estimator of the mean. As noted in section 2.2, $\hat{Y}_{PS}$ will be conditionally biased if the FSU post-stratum totals for the variable of interest, conditional on the number of units in each FSU/post-stratum, follow a model with a non zero intercept. With this in mind, we generated the population in such a way that

$$E\left(y_{ik}\middle|N_{ik}\right)=\alpha_k+\beta N_{ik}+\gamma N_{ik}^2 \qquad (4)$$

where N_{ik} is the number of units in the k^{th} post-stratum for the i^{th} FSU and α_k, β, and γ are constants. Specifically, five post-strata were used with $\alpha_k=100k$ (k=1,...,5), $\beta=10$, and $\gamma=-.05$. In total two thousand FSUs were generated with the total number of units in the i^{th} FSU, say N_i, being a Poisson random variable with mean 10. Then, conditional on N_i, the numbers of units in the five post-strata (i.e., $N_{i1}, N_{i2}, \cdots, N_{i5}$) for the i^{th} FSU were determined using a multinomial distribution with parameters N_i and $p_k=.20$ for $k=1,2,\cdots,5$.

Finally, the value of the variable of interest for the j^{th} unit in the k^{th} post-stratum for the i^{th} FSU was a realization of the random variable

$$y_{ijk}=\alpha_k/N_{ik}+\beta+\gamma N_{ik}+\varepsilon_{1i}+\varepsilon_{2ik}+\varepsilon_{3ijk}N_i$$

where ε_{1i}, ε_{2ik}, and ε_{3ijk} are three independent standardized chi-square (6 d.f.) random variables. This structure implies that $E\left(y_{ik}\middle|N_{ik}\right)$ is given by (4). Furthermore, the values of the variable of interest for units within an FSU are correlated and the correlation depends upon whether the units are in the same post-stratum or not.

A single-stage stratified design was used for the NHIS population with "households" being the FSUs. Ten design strata were used and an approximate 10% simple random sample of households was selected without replacement from each stratum. Each sample consisted of 115 households and each sample household was enumerated completely. A total of 5,000 such samples was selected for the simulation study.

Two-stage stratified sample designs were used for both the CPS and artificial populations. For the CPS population, geographic segments, employed in the original survey and composed of about four neighboring households, were used as FSUs and persons were the second-stage units. In both populations, 100 design strata were created with each stratum having approximately the same number of FSUs and a sample of $m = 2$ FSUs was selected with probability proportional to size from each stratum using the systematic sampling method described by Hansen, Hurwitz, and Madow (1953, p. 343). Thus, 200 FSUs were selected for both populations. Second stage selection was also similar for both populations. For the CPS population a simple random sample of 4 persons was selected without replacement in each sample FSU having $N_i > 4$ and all persons were selected in each sample FSU where $N_i \le 4$. For the artificial population the within FSU sample size was set at 15 rather than 4 which resulted in the complete enumeration of most sample FSUs. A total of 5,000 samples were selected from each of the populations for the simulation study.

In each sample, we computed $\hat{Y}_{HT}$, $\hat{Y}_R$, $\hat{Y}_{PS}$, and two versions of $\hat{Y}_{LR}$. For the first version of the regression estimator, denoted $\hat{Y}_{LR}$(emp) in the tables, **H** was estimated separately from each sample as would be required in practice. Each component of Σ_{12} and Σ_{22} was estimated using the ultimate cluster estimator of covariance, appropriate to the design, as defined in Hansen, et. al. (1953, p.419). The second version, denoted $\hat{Y}_{LR}$(theo) , used the same value of **H** in each sample, which was an estimate more nearly equal to the theoretical value of the **H** matrix. For the CPS and artificial populations, the theoretical **H** matrix was estimated from empirical covariances derived from separate simulation runs of 5,000 samples. For the NHIS population the design was simple enough that a direct theoretical calculation of **H** was done. As the sample of FSUs becomes large, the performance of $\hat{Y}_{LR}$(emp) should approach that of $\hat{Y}_{LR}$(theo). The performance of $\hat{Y}_{LR}$(theo) is, consequently, a gauge of the best that can be expected from the empirical version of the regression estimator for a given sample size.

Table 2 lists unconditional results summarized over all 5,000 samples from each population. Empirical root mean square errors (*rmse*'s) were calculated as $rmse\left(\hat{Y}\right) = \left[\sum_{s=1}^{S} \left(\hat{Y}_s - \bar{Y}\right)^2 \Big/ S \right]^{1/2}$ with S = 5,000 and $\hat{Y}_s$ being one of the estimates of the population mean from sample *s*. In the CPS and artificial populations, results for the Horvitz-Thompson and the ratio estimators were nearly identical so that only the former is shown. Across all samples, the bias of each of the estimators was negligible. As anticipated by the theory, $\hat{\bar{Y}}_{LR}$(theo) was the most precise of the choices, although the largest gain compared to $\hat{\bar{Y}}_{PS}$ was only 4.7% in the artificial population. The need to estimate **H** destabilizes the regression estimator as shown in the results for $\hat{\bar{Y}}_{LR}$(emp). For the NHIS and CPS populations, $\hat{\bar{Y}}_{LR}$(emp) has a larger root *mse* than both $\hat{\bar{Y}}_{LR}$(theo) and $\hat{\bar{Y}}_{PS}$. The most noticeable loss is for the NHIS population where the root *mse* of $\hat{\bar{Y}}_{LR}$(emp) is about 15% larger than that of either $\hat{\bar{Y}}_{LR}$(theo) or $\hat{\bar{Y}}_{PS}$. This result is consistent with the smaller FSU sample size and hence less stable estimate of **H** for the NHIS population.

Figures 1-3 present conditional simulation results. The 5,000 samples were sorted by the theoretical bias factors presented in section 2.2. The sorting was done separately for each of the estimators of the population mean. In the cases of the two regression estimators, which are theoretically unbiased in large samples, the bias factor for $\hat{\bar{Y}}_{PS}$ was used for sorting. The sorted samples were then put into 25 groups of 200 samples each and empirical biases and root *mse*'s were computed within each group. The group results were then plotted versus theoretical bias factors in the figures. The upper sets of points in each figure are the empirical root *mse*'s of the groups, while the lower sets are empirical biases. The two regression estimators are conditionally unbiased as expected. The other estimators, however, have substantial conditional biases that, in the most extreme sets of samples, are important parts of the *mse*'s. For the CPS population, the range of the bias factors for $\hat{\bar{Y}}_{HT}$ is so much larger (-10 to 10) than that of the other estimators that we have omitted $\hat{\bar{Y}}_{HT}$ from the plot for clarity. In the neighborhood of the balance point, $\hat{\bar{\mathbf{N}}} = \bar{\mathbf{N}}$, all estimators perform about the same, but, because of a lack of data at the design stage, we have no control on how close to balance a particular sample may be. The safest choice for controlling conditional bias is, thus, $\hat{\bar{Y}}_{LR}$(emp). This finding is similar to that of Valliant (1990), who noted that, in one-stage, stratified random or systematic sampling, the separate linear regression estimator is a good choice for controlling bias, conditional on the sample mean of an auxiliary variable.

4. DEFECTIVE FRAMES

4.1 The Basic Problem of Defective Frames

In most real world applications not all of the elementary units in the population are included in the sampling frame. In household surveys, it is not unusual for some demographic subgroups, especially minorities, to be poorly covered by the sampling frame. Bailar (1989), for example, notes that in 1985 the sample estimate from the CPS of the total number of Black males, ages 22-24, was only 73% of an independent estimate of the total population of that group. Corresponding percentages for Black males, ages 25-29 and 60-61, were 80% and 76%.

To formalize the discussion of this type of coverage problem, suppose that N_k now refers to the number of elementary units in the frame and that $\dot{N}_k$ is the <u>actual</u> number of population elements in the k^{th} post-stratum. In the discussion below terms with a dot on the top are population values while terms with no dot are frame values. Letting $\dot{Y}_k$ be the aggregate of the y values over all population elements in the k^{th} post-stratum, then it follows that the <u>true population mean</u> is given by

$$\dot{\mu} = \lim_{L \to \infty} \frac{\sum_{k=1}^{K} \dot{Y}_k}{\sum_{k=1}^{K} \dot{N}_k} = \lim_{L \to \infty} \sum_{k=1}^{K} \frac{\dot{N}_k}{\dot{N}} \frac{\dot{Y}_k}{\dot{N}_k} = \sum_{k=1}^{K} \dot{\phi}_k \dot{\mu}_k .$$

Obviously, all four of the estimators of the mean given in section 2 are biased (both conditionally and unconditionally) for $\dot{\mu}$; the additional bias term being given by $\mu - \dot{\mu}$ for all of the estimators. It should be noted that this bias term is $o(1)$ so it will dominate the other bias terms listed in section 2.2 as the number of FSUs increases. There is another even more basic problem; namely, in most cases the individual frame values N_k

are not known so only the ratio estimator is well defined. For example, the Horvitz-Thompson estimator of the mean as defined in section 2 requires N, the total number of units in the frame, but N may be unknown. On the other hand, the $\tilde{N}_k$ (or least the proportions $\tilde{\phi}_k$) may be known from independent sources and hence be available for the purposes of estimator construction. In household surveys, for instance, the $\tilde{N}_k$ may come from intercensal projections of population counts.

Before attempting to construct unbiased estimators for $\tilde{\mu}$ it should be noted that

$$\mu - \tilde{\mu} = \sum_{k=1}^{K} \left(\phi_k - \tilde{\phi}_k\right)\left(\mu_k - \tilde{\mu}_k\right) + \sum_{k=1}^{K} \left(\phi_k - \tilde{\phi}_k\right)\tilde{\mu}_k + \sum_{k=1}^{K} \tilde{\phi}_k\left(\mu_k - \tilde{\mu}_k\right).$$

So, if we assume that for each post-strata the mean of the units in the frame is equal to the true population mean, (i.e. $\mu_k = \tilde{\mu}_k$ for every k) then the bias term reduces to

$$\mu - \tilde{\mu} = \sum_{k=1}^{K} \left(\phi_k - \tilde{\phi}_k\right)\mu_k = \sum_{k=1}^{K} \left(\phi_k - \tilde{\phi}_k\right)\tilde{\mu}_k.$$

This is very strong (and also very expedient) assumption; however, addressing the problem of defective frame bias without such a condition is virtually impossible.

4.2 Alternative Estimators

The basic strategy is to construct an estimator for the defective frame bias, $\mu - \tilde{\mu}$, and then subtract this estimator from the estimators studied earlier. Two cases need to be considered:

Case 1. The frame parameters $\left\{\phi_k, 1 \le k \le K\right\}$ are unknown, and

Case 2. The frame parameters $\left\{\phi_k, 1 \le k \le K\right\}$ are known.

Case 1. For this case only the ratio estimator is well defined and the only obvious candidate for an estimator of the bias is

$$\hat{B}_1 = \sum_{k=1}^{K} \left(\frac{\hat{N}_k}{\hat{N}} - \tilde{\phi}_k\right)\frac{\hat{Y}_k}{\hat{N}_{.k}} = \hat{\bar{Y}}_R - \sum_{k=1}^{K} \tilde{\phi}_k \frac{\hat{Y}_k}{\hat{N}_{.k}}.$$

Using the strategy given above, the resulting estimator for $\tilde{\mu}$ is

$$\hat{\bar{Y}}_1 = \hat{\bar{Y}}_R - \hat{B}_1 = \sum_{k=1}^{K} \tilde{\phi}_k \frac{\hat{Y}_k}{\hat{N}_k}.$$

This is the "post-stratified" estimator usually found in practice. It is straightforward to verify the following properties of $\hat{\bar{Y}}_1$:

$$\mathrm{E}\left[\hat{\bar{Y}}_1 \middle| \hat{\mathbf{N}}\right] = \tilde{\mu} + \left[\mathbf{p}'\mathbf{P}\left(\hat{\mathbf{N}} - \mathbf{M}_1\right)\right] + o\left(m^{-1}\right) \text{ where } \mathbf{p}' = \left[\frac{\tilde{\phi}_1}{\phi_1}, \frac{\tilde{\phi}_2}{\phi_2}, \ldots, \frac{\tilde{\phi}_K}{\phi_K}\right]$$

$$\mathrm{var}\left[\hat{\bar{Y}}_1 \middle| \hat{\mathbf{N}}\right] = m^{-1}\left[\mathbf{p}'\mathbf{V}_{\mathbf{c}}\mathbf{p}\right] + o\left(m^{-(3/2)}\right)$$

$$\mathrm{E}\left[\hat{\bar{Y}}_1\right] = \tilde{\mu} + o\left(m^{-1}\right)$$

$$\mathrm{var}\left[\hat{\bar{Y}}_1\right] = m^{-1}\left[\mathbf{p}'\left[\Sigma_{11} - 2\mathbf{D}(\mu_k)\Sigma_{21} + \mathbf{D}(\mu_k)\Sigma_{22}\mathbf{D}(\mu_k)\right]\mathbf{p}\right] + o\left(m^{-(3/2)}\right)$$

The attempt to correct for the defective frame bias is successful in the sense that $\hat{\bar{Y}}_1$ is unconditionally unbiased for $\tilde{\mu}$. However, the conditional bias is still present.

Case 2. For this case it can be verified that the estimator

$$\hat{B}_2 = (\mathbf{1} - \mathbf{p})'\left[\hat{\bar{\mathbf{Y}}} - \Sigma_{12}\Sigma_{22}^{-1}\left(\frac{\hat{\mathbf{N}}}{\hat{N}} - \mathbf{M}_2\right)\right]$$

is approximately, conditionally unbiased for $\mu - \tilde{\mu}$ and, as $\hat{\bar{Y}}_{LR}$ is conditionally unbiased for μ, it follows directly that the estimator

$$\hat{\bar{Y}}_2 = \hat{\bar{Y}}_{LR} - \hat{B}_2 = \mathbf{p}'\left[\hat{\bar{\mathbf{Y}}} - \Sigma_{12}\Sigma_{22}^{-1}\left(\frac{\hat{\mathbf{N}}}{\hat{N}} - \mathbf{M}_2\right)\right]$$

is both conditionally and unconditionally, approximately unbiased for $\tilde{\mu}$. It can also be verified that

$$\mathrm{var}\left[\hat{\bar{Y}}_2 \middle| \hat{\mathbf{N}}\right] = \mathrm{var}\left[\hat{\bar{Y}}_2\right] = m^{-1}\left[\mathbf{p}'\mathbf{V}_{\mathbf{c}}\mathbf{p}\right] .$$

In addition to the problems of the linear regression estimator cited earlier, this estimator is usually not even well defined as the frame parameters $\left\{\phi_k, 1 \le k \le K\right\}$ are rarely, if ever, known when the frame is defective.

5. CONCLUSION

This study has generalized the asymptotic techniques suggested by Robinson (1987) to study the problem of post-stratification from a design-based, conditional point-of-view. An important paper in the conditional study of post-stratification was that of Holt and Smith (1979), one of whose basic premises was that $\hat{\bar{Y}}_{PS}$ is conditionally unbiased. This will be true (at least asymptotically) only if $\mathbf{1}'\left(\mathbf{H}-\mathbf{D}(\mu_k)\right)=\mathbf{0}'$; so, in general, this premise is false. In fact, simple random sampling of elementary units may be one of the few realistic cases where this basic premise is true.

From a conditional point of view the linear regression estimator is preferable among the four studied here. Only the regression estimator is conditionally unbiased. The post-stratified estimator is no better (or worse) than either the Horvitz-Thompson or the ratio estimator; all have conditional bias terms of order $m^{-(1/2)}$. All of the estimators have the same conditional variance to terms of order m^{-1}; furthermore, the conditional variance does not depend on $\hat{\bar{\mathbf{N}}}$, the vector of estimated proportions in the post-strata. Consequently, because of its conditional unbiasedness, the regression estimator has the smallest conditional mean square error.

The Horvitz-Thompson, ratio, and post-stratified estimators are unconditionally unbiased. Although somewhat illogical, one might attempt to make a case for the estimators by comparing their unconditional properties with the conditional properties of the linear regression estimator. But even from this mixed perspective, the $\hat{\bar{Y}}_{LR}(\text{theo})$ estimator is clearly superior to the others. Not only is it conditionally unbiased, but the conditional variance of the linear regression estimator can be no larger than the unconditional variance of any of the other estimators. In large FSU samples, the empirical version of the regression estimator will inherit these good properties of $\hat{\bar{Y}}_{LR}(\text{theo})$ and also perform well.

The problem of a defective frame introduces complications not found otherwise. Each of the estimators of the mean studied here is biased both conditionally and unconditionally. Bias adjustments are possible only under the restrictive assumption that the mean of units within each post-stratum is the same for all population units whether they are included or excluded from the frame.

An area we have not addressed is variance estimation. A design-based variance estimator for the regression estimator can be obtained using the methods of Särndal, Swensson, and Wretman (1989).

ACKNOWLEDGMENT

Any opinions expressed are those of the authors and do not reflect policy of the U.S. Bureau of Labor Statistics.

REFERENCES

BAILAR, B. (1989). Information Needs, Surveys, and Measurement Errors. In *Panel Surveys,* eds D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh. New York: Wiley.

DURBIN, J. (1969). Inferential Aspects of Randomness of Sample Size in Survey Sampling. In *New Developments in Survey Sampling*, N.L. Johnson and H. Smith, eds. New York: Wiley.

ERICSON, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. *Journal of the Royal Statistical Society B***,** 31, 195-233.

FULLER, W. A. (1981). Comment on An Empirical Study of the Ratio Estimator and Estimators of its Variance by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory,*Vol. 1. New York: John Wiley and Sons.

HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association,* 78, 776-796.

HIDIROGLOU, M. and SÄRNDAL, C.-E. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association*, 84, 266-275.

HOLT, D. and SMITH, T.M.F. (1979). Post Stratification. *Journal of the Royal Statistical Society A*, 142, 33-46.

KREWSKI, D. and RAO, J.N.K. (1981). Inference from Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods. *Annals of Statistics*, 9, 1010-1019.

LITTLE, R.J.A. (1991). Post-Stratification: A Modeler's Perspective. *Proceeding of the Section on Survey Methods Research*, Washington: American Statistical Association, in press.

RAO, J.N.K. (1985). Conditional Inference in Survey Sampling. *Survey Methodology*, 11, 15-31.

__________ (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Presented at the Workshop on Uses of Auxiliary Information in Surveys, Statistics Sweden.

RAO, J.N.K. and WU, C.F.J. (1985). Inference from Stratified Samples: Second Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, 80, 620-630.

ROBINSON, J. (1987). Conditioning Ratio Estimates Under Simple Random Sampling. *Journal of the American Statistical Association*, 82, 826-831.

ROYALL, R.M. (1971). Linear Regression Models in Finite Population Sampling Theory. In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, eds. Toronto: Holt, Rinehart, and Winston.

SÄRNDAL C.-E., SWENSSON, B., and WRETMAN, J. (1989). The Weighted Residual Technique for Estimating the Variance of the Finite Population Total. Biometrika, 76, 527-537.

SÄRNDAL C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

VALLIANT, R. (1990). Comparisons of Variance Estimators in Stratified Random and Systematic Sampling. *Journal of Official Statistics*, 6, 115-131.

________ (1993). Post-stratification and Conditional Variance Estimation. *Journal of the American Statistical Association*, 88, In press.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, 3rd ed. London: Griffin.

Table 1. Population size and basic sample design parameters for three simulation studies.

Population	Pop. Size *N*	No. of FSUs *M*	No. of sample FSUs *m*
HIS	2,934	1,100	115
CPS	10,841	2,826	200
Artificial	22,001	2,000	200

Table 2. Simulation results for three populations. 5,000 samples were selected from each population.

Estimator	Rel-bias $\hat{Y}$ (%)	$rmse\left(\hat{Y}\right)$	$00*\left[\frac{rmse\left(\hat{Y}\right)}{rmse\left(\hat{Y}_{PS}\right)}-1\right]$
HIS population			
$\hat{Y}_{HT}$	.12	.141	.05
$\hat{Y}_{R}$	.10	.141	.02
$\hat{Y}_{PS}$	.11	.141	0
$\hat{Y}_{LR}$(emp)	.19	.162	14.71
$\hat{Y}_{LR}$(theo)	.08	.140	-.96
CPS population			
$\hat{Y}_{HT}$	-.01	10.25	15.8
$\hat{Y}_{PS}$	0	8.85	0
$\hat{Y}_{LR}$(emp)	-.03	9.11	3.0
$\hat{Y}_{LR}$(theo)	-.01	8.79	-.6
Artificial population			
$\hat{Y}_{HT}$	.02	2.30	-2.93
$\hat{Y}_{PS}$	.12	2.37	0
$\hat{Y}_{LR}$(emp)	.04	2.31	-2.41
$\hat{Y}_{LR}$(theo)	.02	2.26	-4.70