# POST-STRATIFICATION AND CONDITIONAL VARIANCE ESTIMATION

Richard Valliant U.S. Bureau of Labor Statistics
Room 2126 , 441 G St. NW Washington DC 20212

## 1. INTRODUCTION

In complex large-scale surveys, particularly household surveys, post-stratification is a commonly used technique for improving efficiency of estimators. A clear description of the method and the rationale for its use was given by Holt and Smith (1979) and is paraphrased here. Values of variables for persons may vary by age, race, sex, and other demographic factors that are unavailable for sample design at the individual level. A population census may, however, provide aggregate information on such variables that can be used at the estimation stage. After sample selection, individual units are classified according to the factors and the known total number of units in the $c$th cell, $M_c$, is used as a weight to estimate the cell total for some target variable. The cell estimates are then summed to yield an estimate for the full population.

Because post-stratum identifiers are unavailable at the design stage, the number of sample units selected from each post-stratum is a random variable. Inferences can be made either unconditionally, i.e. across all possible realizations of the post-strata sample sizes, or conditionally given the achieved sample sizes. In a simpler situation than that considered here, Durbin (1969) maintained, on grounds of common sense and the ancillarity of the achieved sample size, that conditioning was appropriate. In the case of post-stratification in conjunction with simple random sampling of units, Holt and Smith (1979) argue strongly that inferences should be conditioned on the achieved post-stratum sample sizes.

Although conditioning is, in principle, a desirable thing to do, a design-based conditional theory for complex surveys may be intractable, as noted by Rao (1985). A useful alternative is the prediction or superpopulation approach which is applied in this paper to inference from post-stratified samples. We will concentrate especially on the properties of two commonly used variance estimators to determine whether they estimate the conditional variance of the post-stratified estimator of a finite population total.

Section 2 introduces notation, a superpopulation model that will be used to study properties of various estimators, and a class of estimators which will be used as the starting point for post-stratification estimation. Section 3 discusses the model bias and variance of estimators of the total while sections 4 and 5 cover the linearization and balanced repeated replication variance estimators. In section 6 we present the results of a simulation study using data from the U.S. Current Population Survey and the last section gives concluding remarks.

## 2. NOTATION AND MODEL

The population of units is divided into H design strata with stratum h containing $N_h$ clusters. Cluster $(hi)$ contains $M_{hi}$ units with the total number of units in stratum $h$ being $M_h = \sum_{i=1}^{N_h} M_{hi}$ and the total in the population being $M = \sum_{h=1}^{H} M_h$. A two-stage sample is selected from each stratum consisting of $n_h \geq 2$ sample clusters and a subsample of $m_{hi}$ sample units within sample cluster $(hi)$. The total number of clusters in the sample is $n = \sum_h n_h$. The set of sample clusters from stratum $h$ is denoted by $s_h$ and the subsample of units within sample cluster $(hi)$ by $s_{hi}$.

Associated with each unit in the population is a random variable $y_{hij}$ whose finite population total is $T = \sum_h \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} y_{hij}$. Each unit is also a member of a class or post-stratum indexed by $c$. Each post-stratum can cut across the design strata and the set of all population units in post-stratum $c$ is denoted by $S_c$. The total number of units in post-stratum $c$ is $M_c = \sum_h \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_c \delta_{hijc}$ where $\delta_{hijc} = 1$ if unit $(hij)$ is in post-stratum $c$ and is 0 if not. We assume that the post-stratum sizes $M_c$ are known. Our goal here will be to study the properties of estimators under the following superpopulation model:

$$E(y_{hij}) = \mu_c$$

$$\operatorname{cov}(y_{hij}, y_{h'i'j'}) = \begin{cases} \sigma^2_{hic} & h = h', i = i', j = j', \\ & (hij) \in S_c \\ \sigma^2_{hic}\rho_{hic} & h = h', i = i', j \neq j', \\ & (hij) \in S_c, (h'i'j') \in S_c \\ \tau_{hicc'} & h = h', i = i', j \neq j', \\ & (hij) \in S_c, (h'i'j') \in S_{c'} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In addition to being uncorrelated, we also assume that the y's associated with units in different clusters are independent. The model assumes that units in a post-stratum have a common mean $\mu_c$ and are correlated within a cluster. The size of the covariances $\sigma^2_{hic}\rho_{hic}$ and $\tau_{hicc'}$ are allowed to vary among the clusters and also depend on whether or not units are in the same post-stratum. The variance specification $\sigma^2_{hic}$ is quite general, depending on the design stratum, cluster, and post-stratum associated with the unit.

The general type of estimator of $T$ that we will consider has the form

$$\hat{T} = \sum_h \sum_{i \in s_h} \gamma_{hi} \hat{T}_{hi} \qquad (2)$$

where $\gamma_{hi}$ is a coefficient that does not depend on the y's, $\hat{T}_{hi} = M_{hi}\bar{y}_{hi}$, and $\bar{y}_{hi} = \sum_{j \in s_{hi}} y_{hij}/m_{hi}$. In common survey practice, the set of $\gamma_{hi}$ is selected to produce a design-unbiased or design-consistent estimator of the total under the particular probability sampling design being used. Alternatively, estimator (2) can be written as

$$\hat{T} = \sum_h \sum_{i \in s_h} \sum_c K_{hic}\bar{y}_{hic} \qquad (3)$$

where $K_{hic} = \gamma_{hi} M_{hi} m_{hic}/m_{hi}$, $m_{hic}$ is the number of sample units in sample cluster $(hi)$ that are part of post-stratum $c$, and $\bar{y}_{hic} = \sum_{j \in s_{hi}} y_{hij}\delta_{hijc}/m_{hic}$. If $m_{hic} = 0$, then define $\bar{y}_{hic} = 0$. There are a variety of estimators, both from probability sampling theory and superpopulation theory, that fall in this class. Six examples are given in Valliant (1987) and include types of separate ratio and regression estimators with $M_{hi}$ used as the auxiliary variable. Also included in class (2) is the Horvitz-Thompson estimator when clusters are selected with probabilities proportional to $M_{hi}$ and units within clusters are selected with equal probability in which case $\gamma_{hi} = M_h/(n_h M_{hi})$. Note that, as discussed in section 3, the estimators defined by (2) are not necessarily model-unbiased under (1).

Next, we turn to the definition of the post-stratified estimator of the total. The usual design-based estimator of $M_c$ in class (2) is found by using $\delta_{hijc}$ in place of $y_{hij}$ in (3) and omitting the sum over $c$, which gives

$$\hat{M}_c = \sum_h \sum_{i \in s_h} K_{hic}.$$

The post-stratified estimator of the total $T$ is then defined as

$$\hat{T}_{ps} = \sum_c \hat{R}_c \hat{T}_c \qquad (4)$$

where $\hat{R}_c = M_c/\hat{M}_c$ and $\hat{T}_c = \sum_h \sum_{i \in s_h} K_{hic}\bar{y}_{hic}$. With this notation the general estimator (3) can also be written as $\hat{T} = \sum_c \hat{T}_c$.

## 3. Bias and Variance of Estimators of the Total

The model bias under (1) of the unadjusted estimator $\hat{T}$ is

$$E(\hat{T} - T) = \sum_c \mu_c(\hat{M}_c - M_c).$$

Estimators in class (2) are model unbiased if $\hat{M}_c = M_c$ a condition which in general does not hold for a particular sample but may be true in expectation across all samples that a particular design can produce. On the other hand, the post-stratified estimator $\hat{T}_{ps}$ is model-unbiased under (1), as is easily verified.

The prediction variance of the post-stratified estimator is defined as $\mathrm{var}(\hat{T}_{ps} - T)$. Under some reasonable assumptions, similar to those given in Royall (1986) or Valliant (1987), on how certain population and sample quantities grow as $H \to \infty$, we have

$$\mathrm{var}(\hat{T}_{ps} - T) \approx \mathrm{var}(\hat{T}_{ps}) \qquad (5)$$

where $\approx$ denotes "asymptotically equivalent to." Details are sketched in Valliant (1991). Consequently, we will concentrate on the estimation of $\mathrm{var}(\hat{T}_{ps})$.

In order to compute the variance, it is convenient to write the post-stratified estimator as

$$\hat{T}_{ps} = \sum_h \hat{\mathbf{R}}' \hat{\mathbf{T}}_h$$

where

$$\hat{\mathbf{R}} = (\hat{R}_1, \ldots, \hat{R}_C)',$$

$$\hat{\mathbf{T}}_h = (\mathbf{K}'_{h1}\bar{\mathbf{y}}_{h1}, \ldots, \mathbf{K}'_{hC}\bar{\mathbf{y}}_{hC})',$$

$$\mathbf{K}'_{hc} = (K_{h1c}, \ldots, K_{hn_hc})', \text{ and}$$

$$\bar{\mathbf{y}}_{hc} = (\bar{y}_{h1c}, \ldots, \bar{y}_{hn_hc})'.$$

The variance can be found as

$$\mathrm{var}(\hat{T}_{ps}) = \sum_h \hat{\mathbf{R}}' \mathbf{K}'_h \mathbf{V}_h \mathbf{K}_h \hat{\mathbf{R}} \qquad (6)$$

where $\mathbf{K}'_h$ is the $C \times n_h C$ matrix whose $c^{\text{th}}$ row is $(\mathbf{0}'_{n_h}, \ldots \mathbf{0}'_{n_h}, \mathbf{K}'_{hc}, \mathbf{0}'_{n_h}, \ldots \mathbf{0}'_{n_h})$, i.e. $\mathbf{K}'_{hc}$ is preceded by c-1

zero row vectors of length $n_h$ and followed by $C$-$c$ such zero vectors. The matrix $\mathbf{V}_h$ is defined as

$$\mathbf{V}_h = \begin{bmatrix} \mathbf{V}_{h1} & \mathbf{D}_{sh12} & \cdots & \mathbf{D}_{sh1C} \\ \mathbf{D}_{sh21} & \mathbf{V}_{h2} & & \\ \vdots & & \ddots & \\ \mathbf{D}_{shC1} & & & \mathbf{V}_{hC} \end{bmatrix}$$

with $\mathbf{V}_{hc} = \mathrm{diag}(v_{hic})_{n_h \times n_h}$ and $\mathbf{D}_{shcc'} = \mathrm{diag}(\tau_{hicc'})_{n_h \times n_h}$ where $i \in s_h$ for both $\mathbf{V}_{hc}$ and $\mathbf{D}_{shcc'}$. A key point is that, although the factors $\hat{R}_c$ may be random with respect to the sample design, they are constant with respect to model (1), so that (6) is a variance conditional on the values of $\hat{R}_c$.

The variance of the unadjusted estimator $\hat{T}$ can be found by minor modification of the above arguments. Because $\hat{T} = \sum_h \hat{\mathbf{T}}_h$, i.e. the value of $\hat{T}_{ps}$ when $\hat{R}_c = 1$ for all $c$, we have

$$\mathrm{var}(\hat{T}) = \sum_h \mathbf{1}_C' \mathbf{K}_h' \mathbf{V}_h \mathbf{K}_h \mathbf{1}_C' \qquad (7)$$

where $\mathbf{1}_C$ is a vector of $C$ 1's. Note that, if the sample design is such that the post-stratum factors $\hat{R}_c$ each converge to 1, then $\mathrm{var}(\hat{T}_{ps})$ and $\mathrm{var}(\hat{T})$ are about the same in large samples.

## 4. A Linearization Variance Estimator

Linearization or Taylor series variance estimators for post-stratified estimators are discussed for general sample designs by Fuller and Sullivan (1987), Rao (1985), and Williams (1962). An application to a complex survey design is given in Parsons and Casady (1985). Our interest here is in how a linearization estimator, derived from design-based arguments, performs as an estimator of the approximate conditional variance given by (6). For clarity and completeness we will sketch the derivation of the estimator for the class of post-stratified estimators studied here. In a design-based analysis, the product $\hat{R}_c\hat{T}_c$ is expanded about the point $(M_c, T_c)$ where $T_c$ is the finite population total for post-stratum $c$. The usual first-order Taylor approximation to $\hat{R}_c\hat{T}_c$ is

$$\hat{R}_c\hat{T}_c \cong T_c + \hat{T}_c - \frac{T_c}{M_c}\hat{M}_c.$$

From this expression it follows that

$$\hat{T}_{ps} - T \cong \sum_h \sum_{i \in s_h} \tilde{d}_{hi}$$

where $\tilde{d}_{hi} = \sum_{j \in s_{hi}} \sum_c \gamma_{hi} M_{hi} \delta_{hijc} \left[ y_{hij} - (T_c/M_c) \right] / m_{hi}$. For computations, the usual procedure is to substitute estimators for the unknown quantities in $\tilde{d}_{hi}$ producing

$$d_{hi} = \sum_{j \in s_{hi}} \sum_c \gamma_{hi} M_{hi} \delta_{hijc} (y_{hij} - \hat{\mu}_c) / m_{hi}$$
$$= \sum_c K_{hic} (\bar{y}_{hic} - \hat{\mu}_c)$$

where $\hat{\mu}_c = \hat{T}_c / \hat{M}_c$. The linearization variance estimator, including an *ad hoc* finite population correction factor, is then defined as

$$v_L(\hat{T}_{ps}) = \sum_h \frac{n_h}{n_h - 1}(1 - f_h) \sum_{i \in s_h} (d_{hi} - \bar{d}_h)^2 \qquad (8)$$

where $f_h = n_h / N_h$ and $\bar{d}_h = \sum_{i \in s_h} d_{hi} / n_h$.

In order to determine whether the general linearization estimator (8) estimates the conditional variance (6), we examine its large sample behavior. As shown in Valliant (1991)

$$\frac{n}{M^2}\left[ v_L(\hat{T}_{ps}) - \sum_h \mathbf{1}_C' \mathbf{K}_h' \mathbf{V}_h \mathbf{K}_h \mathbf{1}_C \right] \xrightarrow{p} 0.$$

Thus, the linearization estimator $v_L$ actually estimates $\mathrm{var}(\hat{T})$ given by (7) rather than $\mathrm{var}(\hat{T}_{ps})$ in (6). In large samples the linearization estimator differs from $\mathrm{var}(\hat{T}_{ps})$ by a factor that depends on how different the adjustment factors $\hat{\mathbf{R}}$ are from $\mathbf{1}_C$.

This conditional bias can be eliminated by using the adjusted deviate

$$d_{hi}^* = \sum_{j \in s_{hi}} \sum_c \hat{R}_c \gamma_{hi} M_{hi} \delta_{hijc} (y_{hij} - \hat{\mu}_c) / m_{hi}$$
$$= \sum_c \hat{R}_c K_{hic} (\bar{y}_{hic} - \hat{\mu}_c)$$

in the linearization estimator. For later reference define the adjusted linearization estimator using $d_{hi}^*$ as $v_L^*(\hat{T}_{ps})$ and note that it can be written as

$$v_L^*(\hat{T}_{ps}) = \sum_h \frac{n_h}{n_h - 1}(1 - f_h) \sum_{i \in s_h} \left[ \sum_c \hat{R}_c (d_{hic} - \bar{d}_{hc}) \right]^2.$$

Proof that this adjusted variance estimator is consistent for $\mathrm{var}(\hat{T}_{ps} - T)$ is given in Valliant (1991).

## 5. A Balanced Repeated Replication Variance Estimator

Balanced repeated replication (*BRR*) or balanced half-sample variance estimators, proposed

by McCarthy (1969), are often used in complex surveys because of their generality and the ease with which they can be programmed. Suppose $n_h = 2$ in all strata. A set of $J$ half-samples is defined by the indicators

$$\varsigma_{hi\alpha} = \begin{cases} 1 & \text{if unit } i \text{ is in half-sample } \alpha \\ 0 & \text{if not} \end{cases}$$

for $i$=1,2 and $\alpha$=1,..., $J$. Based on the $\varsigma_{hi\alpha}$, define

$$\varsigma_h^{(\alpha)} = 2\varsigma_{h1\alpha} - 1 = \begin{cases} 1 & \text{if unit } h1 \text{ is in half-sample } \alpha \\ -1 & \text{if unit } h2 \text{ is in half-sample } \alpha. \end{cases}$$

Note also that $-\varsigma_h^{(\alpha)} = 2\varsigma_{h2\alpha} - 1$. A set of half-samples is orthogonally balanced if $\sum_{\alpha=1}^{J} \varsigma_h^{\alpha} = \sum_{\alpha=1}^{J} \varsigma_h^{\alpha}\varsigma_{h'}^{\alpha} = 0\,(h' \neq h)$ with a minimal set of half-samples having $H+1 \le J \le H+4$. One of the choices of balanced half-sample variance estimators is

$$v_{BRR}\left(\hat{T}_{ps}\right) = \sum_{\alpha=1}^{J}\left(\hat{T}_{ps}^{(\alpha)} - \hat{T}_{ps}\right)^2 / J \qquad (9)$$

where $\hat{T}_{ps}^{(\alpha)} = \sum_c \hat{R}_c^{(\alpha)}\hat{T}_c^{(\alpha)}$ with $\hat{R}_c^{(\alpha)}$ being the post-stratum $c$ adjustment factor and $\hat{T}_c^{(\alpha)}$ being the estimated post-stratum total based on half-sample $\alpha$, both of which are defined explicitly below.

In applying the *BRR* method, practitioners often repeat each step of the estimation or weighting process, including post-stratification adjustments, for each half-sample. The intuition behind such repetition is that the variance estimator will then incorporate all sources of variability. The goal here is the estimation of a conditional variance. This raises the question of whether, to achieve that goal, the post-stratification factors $\hat{\mathbf{R}}$ should be recomputed for each half-sample or whether the full-sample factors should be used for each half-sample.

First, consider the case in which the factors are recomputed from each half-sample and define $\hat{R}_c^{(\alpha)} = M_c / \hat{M}_c^{(\alpha)}$ to be the factor and $\hat{T}_c^{(\alpha)}$ to be the estimated total for post-stratum $c$ based on half-sample $\alpha$. In particular,

$$\hat{T}_c^{(\alpha)} = \sum_h \sum_{i \in s_h} (2\varsigma_{hi\alpha} - 1) K_{hic}\bar{y}_{hic} \text{ and}$$

$$\hat{M}_c^{(\alpha)} = \sum_h \sum_{i \in s_h} (2\varsigma_{hi\alpha} - 1) K_{hic}.$$

Next, expand $\hat{R}_c^{(\alpha)}\hat{T}_c^{(\alpha)}$ around the full sample estimates $\hat{R}_c$ and $\hat{T}_c$ to obtain the approximation

$$\hat{R}_c^{(\alpha)}\hat{T}_c^{(\alpha)} - \hat{R}_c\hat{T}_c \cong \hat{R}_c\left(\hat{T}_c^{(\alpha)} - \hat{T}_c\right) - \hat{R}_c\hat{\mu}_c\left(\hat{M}_c^{(\alpha)} - \hat{M}_c\right)$$

$$= \hat{R}_c\left[\sum_h \varsigma_h^{(\alpha)}\left(\Delta_{yhc} - \hat{\mu}_c\Delta_{Khc}\right)\right] \qquad (10)$$

where

$$\Delta_{yhc} = K_{h1c}\bar{y}_{h1c} - K_{h2c}\bar{y}_{h2c} \text{ and}$$

$$\Delta_{Khc} = K_{h1c} - K_{h2c}.$$

The variance estimator (9) can then be approximated as

$$v_{BRR}\left(\hat{T}_{ps}\right) \cong \sum_{\alpha=1}^{J}\left[\sum_c \hat{R}_c \sum_h \varsigma_h^{(\alpha)} z_{hc}\right]^2$$

where $z_{hc} = \Delta_{yhc} - \hat{\mu}_c\Delta_{Khc}$. Squaring out the term in brackets and using the fact that $\varsigma_h^{(\alpha)2} = 1$ and the orthogonality of $\varsigma_h^{(\alpha)}$ and $\varsigma_{h'}^{(\alpha)}$ $(h \neq h')$, lead to

$$v_{BRR}\left(\hat{T}_{ps}\right) \cong \sum_h\left(\sum_c \hat{R}_c z_{hc}\right)^2. \qquad (11)$$

Squaring out the right-hand side of (11) and noting that

$$z_{hc}z_{hc'} = \frac{n_h}{n_h - 1}\sum_{i \in s_h}\left(d_{hic} - \bar{d}_{hc}\right)\left(d_{hic'} - \bar{d}_{hc'}\right)$$

when $n_h = 2$, it follows that, aside from the factor $1 - f_h$, $v_{BRR}\left(\hat{T}_{ps}\right)$ is approximately equal to the adjusted linearization estimator in section 4. Consequently, the *BRR* estimator does appropriately estimate the conditional variance when the number of strata is large and when the post-stratification factors are recomputed for each half-sample.

Suppose, alternatively, that the full-sample factors are used for each half-sample, and denote the resulting estimator as $v_{BRR}^{*}\left(\hat{T}_{ps}\right)$. Expression (10) then becomes

$$\hat{R}_c^{(\alpha)}\hat{T}_c^{(\alpha)} - \hat{R}_c\hat{T}_c = \hat{R}_c\left(\hat{T}_c^{(\alpha)} - \hat{T}_c\right) = \hat{R}_c\sum_h \varsigma_h^{(\alpha)}\Delta_{yhc}$$

and the term $z_{hc}$ in (11) reduces to $z_{hc} = \Delta_{yhc}$. By direct calculation the expectation of approximation (11) is

$$E\left[\sum_h\left(\sum_c \hat{R}_c\Delta_{yhc}\right)^2\right] = \text{var}\left(\hat{T}_{ps}\right) + \sum_h \mu' \mathbf{V}_{Kh}\mu$$

where $\mu' = (\mu_1, \ldots, \mu_C)$ and $\mathbf{V}_{Kh}$ is a $C \times C$ matrix with the $(cc')^{th}$ element equal to

$$\sum_{i \in s_h}\left(K_{hic} - \bar{K}_{hc}\right)\left(K_{hic'} - \bar{K}_{hc'}\right)$$

where $\bar{K}_{hc} = \sum_{i \in s_h} K_{hic}/n_h$. Because $\mathbf{V}_{Kh}$ is a type of covariance matrix, it is positive semi-definite. As a result, using the full sample post-stratification factors in each replicate can lead to an overestimate of the

variance of $\hat{T}_{ps}$. As will be illustrated in the empirical study in section 6, the overestimation can be severe.

## 6. A Simulation Study

The preceding theory was tested in a simulation study using a population of 10,841 persons who were included in the September 1988 Current Population Survey (CPS). The variables used in the study were weekly wages and hours worked per week for each person. The study population contained 2,826 geographic segments. The segments were those used in the CPS with each being composed of about four neighboring households. Eight post-strata were formed on the basis of age, race, and sex using tabulations of weekly wages on the full population. Table 1 shows the age/race/sex categories which were assigned to each post-stratum, and Table 2 gives the means per person of weekly income and hours worked per week in each post-stratum. As is apparent from Table 2, the means differ considerably among the post-strata, especially for weekly wages.

A two-stage stratified sample design was used in which segments were selected as the first-stage units and persons as the second-stage units. Two sets of 10,000 samples were selected. For the first set, 100 sample segments were selected with probabilities proportional to the number of persons in each segment. For the second set, 200 segments were sampled. In both cases, strata were created to have about the same total number of households and $n_h = 2$ sample segments were selected per stratum. A simple random sample of 4 persons was selected without replacement in each segment having $M_{hi} > 4$. In cases having $M_{hi} \le 4$, all persons in the sample cluster were selected.

In each sample, we computed the Horvitz-Thompson estimator $\hat{T}_{HT}$ (which is a special case of the general unadjusted estimator defined in section 2), the post-stratified estimator $\hat{T}_{ps}$, and the four variance estimators $v_L$, $v_L^*$, $v_{BRR}$, and $v_{BRR}^*$. For the two *BRR* estimators, the half-sample total $\hat{T}_c^{(\alpha)}$ was computed as $\hat{T}_c^{(\alpha)} = \sum_h \sum_{i \in s_h} \sqrt{1-f_h}\,(2\varsigma_{hi\alpha} - 1) K_{hic} \bar{y}_{hic}$ which has the effect of inserting finite population correction factors for each stratum in the approximation given by (11). Table 3 presents results summarized over all 10,000 samples. Empirical mean square errors (*mse*'s) were calculated as $mse(\hat{T}) = \sum_{s=1}^{S} (\hat{T}_s - T)^2 / S$ with $S$ = 10,000 and $\hat{T}$ being either $\hat{T}_{HT}$ or $\hat{T}_{ps}$. Average variance estimates across the samples were computed as $\bar{v} = \sum_{s=1}^{S} v_s / S$ where $v_s$ is one of the five variance estimates considered. The table reports the ratios $\sqrt{\bar{v}/mse(\hat{T})}$.

As anticipated by the theory in section 4 the linearization variance estimator $v_L$ is more nearly an estimate of the *mse* of the Horvitz-Thompson estimator $\hat{T}_{HT}$ than of the *mse* of $\hat{T}_{ps}$. The adjusted linearization estimator $v_L^*$, on the other hand, is approximately unbiased for $mse(\hat{T}_{ps})$, as is the jackknife. Of the two *BRR* estimates, the root of the average $v_{BRR}$ performs well while the adjusted choice $v_{BRR}^*$ is a serious overestimate as predicted by the theory in section 5. The estimate $v_{BRR}^*$ is also much more variable than either $v_L^*$ or $v_{BRR}$, as shown in the lower part of Table 3.

## 7. Conclusion

Post-stratification is an important estimation tool in sample surveys. Though often thought of as a variance reduction technique, the method also has a role in reducing the conditional bias of the estimator of a total, as illustrated here. The usual linearization variance estimator for the post-stratified total $\hat{T}_{ps}$ actually estimates an unconditional variance as shown here both theoretically and empirically. This deficiency is easily remedied by a simple adjustment which parallels one that can be made for the case of the ratio estimator. Standard application of the *BRR* and jackknife variance estimators does, on the other hand, produce conditionally consistent estimators. An operational question that is sometimes raised in connection with replication estimators is whether to recompute the post-stratification factors for each replicate or to use the full sample factors in each replicate estimate. Judging from the theoretical and empirical results for *BRR* reported here, recalculation for each replicate is by far the preferable course, leading to a variance estimator that is more nearly unbiased and more stable.

## References


Durbin, J. (1969), "Inferential Aspects of the Randomness of Sample Size in Survey Sampling," in *New Developments in Survey Sampling*, N.L. Johnson and H. Smith, Jr. eds, Wiley-Interscience, 629-651.

Fuller, W.A. and Sullivan, G. (1987), "Gamma Post Stratification," technical report J.S.A. 87-1 for the U.S. Bureau of the Census.

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

McCarthy, P.J. (1969), "Pseudo-replication: Half-Samples," *Review of the International Statistical Institute*, **37**, 239-264.

Parsons, V.L. and Casady, R.J. (1985), "Variance Estimation and the Redesigned National Health Interview Survey, *Proceedings of the Section on Survey Research Methodology*, American Statistical Association, 406-411.

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

Royall, R.M. (1986), "The Prediction Approach to Robust Variance Estimation in Two-Stage Cluster Sampling," *Journal of the American Statistical Association*, **81**, 119-123.

Valliant, R. (1987), "Generalized Variance Functions in Stratified Two-Stage Sampling," *Journal of the American Statistical Association*, **82**, 499-508.

——— (1991), "Post-stratification and Conditional Variance Estimation," *BLS* technical report.

Williams, W.H. (1962), "The Variance of an Estimator with Post-Stratified Weighting," *Journal of the American Statistical Association*, **57**, 622-627.

*Table* 1. Assignment of age/race/sex categories to post-strata. Numbers in cells are post-stratum identification numbers (1 to 8).

| | Non-Black | | Black | |
|---|---|---|---|---|
| Age | Male | Female | Male | Female |
| 19 & under | 1 | 1 | 1 | 1 |
| 20-24 | 2 | 3 | 3 | 3 |
| 25-34 | 5 | 6 | 4 | 4 |
| 35-64 | 7 | 8 | 4 | 4 |
| 65 & over | 2 | 3 | 3 | 1 |

*Table* 2. Means per person in each of the eight post-strata for weekly wages and hours worked per week.

| Post-stratum | No. of persons $M_c$ | Weekly wages | Hours worked |
|---|---|---|---|
| 1 | 815 | 111.1 | 23.7 |
| 2 | 691 | 278.7 | 37.9 |
| 3 | 829 | 221.7 | 34.7 |
| 4 | 955 | 349.7 | 38.8 |
| 5 | 1,543 | 455.9 | 43.5 |
| 6 | 1,262 | 319.1 | 37.5 |
| 7 | 2,541 | 554.2 | 43.1 |
| 8 | 2,205 | 326.9 | 36.4 |
| Total | 10,841 | 372.3 | 38.3 |

*Table 3*. Summary results over two sets of 10,000 two-stage stratified samples of 100 and 200 segments each.

| | Weekly wages | | Hours worked | |
|---|---|---|---|---|
| Summary quantity | n=100 | n=200 | n=100 | n=200 |
| Empirical $\sqrt{mse}(+10^3)$ | | | | |
| $\hat{T}_{HT}$ | 156.1 | 111.1 | 6.8 | 4.9 |
| $\hat{T}_{ps}$ | 138.0 | 96.0 | 6.2 | 4.3 |
| $\sqrt{\bar{v}_L / mse(\hat{T}_{HT})}$ | 1.034 | 1.002 | 1.031 | .993 |
| [Avg. var. est./mse($\hat{T}_{ps}$)]$^{1/2}$ | | | | |
| $v_L$ | 1.170 | 1.159 | 1.122 | 1.110 |
| $v_L^*$ | 1.004 | 1.013 | .999 | .998 |
| $v_{BRR}$ | 1.061 | .993 | 1.028 | .982 |
| $v_{BRR}^*$ | 1.229 | 1.133 | 1.277 | 1.160 |
| $v_J$ | 1.006 | 1.014 | 1.002 | .998 |
| Std. dev. of var. est.$(+10^3)$ | | | | |
| $v_L$ | 5616 | 1935 | 14.7 | 5.1 |
| $v_L^*$ | 4290 | 1500 | 12.2 | 4.2 |
| $v_{BRR}$ | 4941 | 1455 | 13.1 | 4.1 |
| $v_{BRR}^*$ | 6900 | 2022 | 20.9 | 6.2 |
| $v_J$ | 4326 | 1503 | 12.3 | 4.2 |