Efficacy of Poststratification in Complex Sample Designs

Ismael Flores Cervantes[1] and J. Michael Brick[1]
[1]Westat, 1600 Research Blvd, Rockville, Maryland 20850

Abstract
Poststratification is a calibration estimation method that is often used to reduce the variance of the estimates and to reduce bias due to noncoverage or nonresponse. In this paper we examine the efficiency of poststratification in the full response and coverage situation. Virtually all results on the efficiency of poststratification in the literature assume a simple random sampling. We expand this and look at the efficiency in one complex design, a disproportionate stratified random sample. We provide an expression based on the coefficient of determination R^2 to assess the reduction or increase of variance due to poststratification in this type of sample design.

Keywords: Calibration, stratification, poststratification

1. Poststratification

Poststratification is a method of estimation that is very popular among survey practitioners. The main motivation for its use is variance reduction although it has been also used to adjust for nonresponse and noncoverage errors (Kott, 2006). Poststratification is described in Holt and Smith (1979), for example, where after sample selection sampled units are classified into groups and the known total number of units in the population is used to estimate the group total for some variable of interest. The group totals are summed to produce an estimate for the whole population. Typically these groups or poststrata are formed so they contain at least a minimum number of sampled units.

Poststratification or stratification after sampling is used because the information for the classification of the sampling units is not available prior data collection or is very expensive to use when creating sampling strata. The benefits of poststratification as reported in the literature are similar to those from stratification and proportional allocation for a design when these sampling methods are not initially used to select the sample.

Although properties of poststratification and inferences of poststratified estimates have been studied extensively, except for few cases (Williams, 1962), its focus has been mainly on the simple random sample design. Furthermore, the same literature assumes 100 percentage coverage and response (i.e., no biases), and large sample sizes.

The objectives of this paper are to find expressions to evaluate the gains of poststratification in sample designs other than simple random sampling, in particular for stratified simple random designs, and for estimates other than totals such as means. Several authors have suggested that poststratification is disappointing in that it does not improve the efficiency of the estimates very much (e.g., Hartley, 1962; Holt and Smith,

1979). As part of the research we attempt to determine the conditions where there are gains in variance reduction due to poststratification.

1.1 Poststratification Estimator

Poststratification is a method from the class of estimators called calibration estimators (Deville and Särndal, 1992). The poststratified estimator for a total is computed as

$$\hat{y}_{ps} = \sum_g N_g \hat{N}_g^{-1} \hat{y}_g = \sum_g N_g \hat{N}_g^{-1} \sum_{k \in s_g} d_k y_k \qquad (1)$$

where d_k is the inverse of the probability of selection , $\mathbf{N}_G = \left(N_1,...,N_g,...,N_G\right)'$ is a vector of known population totals that define the G poststrata, $\hat{N}_g = \sum_{k \in s_g} d_k$ is the sample estimate of the total N_g, and N is the total population size such that $N = \sum_g N_g$.

1.2 Variance of the Poststratified Estimator

In most textbooks the only estimate of variance for a poststratified estimator explicitly stated is for totals from simple random sampling designs. For example, Cochran (1977) shows that the variance can be computed as

$$V\left(\hat{y}_{ps}\right) \approx \left(1-f\right)\frac{N^2}{n}\sum_g W_g S_g^2 + \frac{N^2}{n}\sum_g \left(1-W_g\right)S_g^2 \qquad (2)$$

where $S_g^2 = \left(N_g - 1\right)^{-1}\sum_{k \in s_g}\left(y_k - \overline{Y}_g\right)$ is the population variance in poststrata g, $g = 1,...,G$. Expression (1) has two components (Cochran 1977, Kish 1995; Thompson 1992); the first component corresponds to the value of $V\left(\hat{y}_{st}\right)$ for a design with a sample of size n proportionally allocated to G strata (i.e., $n_g = nW_g$ where $W_g = N_g / N$). The second term, generally small when the sample size n_g is moderately large, reflects the increase in variance of the estimate due to variation of the sample in the poststrata. That is n_g is random and only in expectation is distributed proportionally among the poststrata. The second term of the variance is derived noticing that n_g follows a hypergeometric distribution with parameters $\left(n, N_g, N\right)$, where $n = \sum_g n_g$, and using the Taylor Series' approximation of $E\left(n_g^{-1}\right)$[1]. The order of this tem is $O\left(1/n^2\right)$.

Although Cochran (1977) and Särndal, Swensson, and Wretman, (1992) mention the use of the poststratified estimator in a stratified design where the strata are created using a variable other than the poststrata, they do not provide an expression of the variance of $\hat{y}_{ps}$ in this situation.

[1] In Cochran (1977), the binomial approximation to the hypegeometric distribution of n_g is used to derive the expression of the second term of the variance (2). The difference is the inclusion of the factor 1-n/N in the formula.

1.3 Poststratification as a Regression Estimator

As mentioned before, poststratification can be studied using different approaches. One of these approaches is regression theory applied to survey sampling (Fuller, 2009, Särndal, et al., 1992). Regression estimators incorporate supplementary (or auxiliary) information at the estimation stage to increase the precision of the estimator. This class of estimators uses a model for the relationship between the variable of interest and the auxiliary variables. Estimators from this class of estimators are model assisted and their efficiency compared with unadjusted estimators depends on the goodness of fit of the regression.

The form of the regression estimator is

$$\hat{y}_{gr} = \left(\sum \mathbf{x}_k\right)' \hat{\mathbf{B}} \sum_{k \in s} d_k e_k \,, \tag{3}$$

where $\mathbf{x}_k$ is the vector of known auxiliary variables, $\hat{\mathbf{B}} = \left(\sum d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_{k \in s} d_k \mathbf{x}_k y_k$, and $e_k = y_k - \mathbf{x}_k \hat{\mathbf{B}}$ is the difference between the predicted value and the observed value in the sample called the residual.

The expression of the approximate variance of the regression estimator is computed using the general formula with the expansion of the residuals as

$$AV\left(\hat{y}_{gr}\right) = \sum_k \sum_l \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right)\left(y_k - \mathbf{x}_k \mathbf{B}\right)\left(y_l - \mathbf{x}_l \mathbf{B}\right), \tag{4}$$

where π_k is the inclusion probability of element k, π_{kl} is the probability that both elements k and l are included in the sample, and $\mathbf{B} = \left(\sum \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum \mathbf{x}_k y_k$.

The poststratification estimator is a special case of the regression estimator with a model that partitions the population in subgroups g (g = 1, ..., G). The model assumes a group mean where elements within the same group share the same mean and variance (i.e., $E\left(y_k\right) = \beta_g$ and $V\left(y_k\right) = \sigma_g^2$). In particular, the auxiliary information matrix $\mathbf{x}$ consist of G - 1 indicators or dummy variables (values 1 or 0) that indicate the group membership of the elements y_k's.

Särndal et al. (1992) provide an expression for the approximate variance of the poststratification estimator under simple random sample (SRS) that is equal to (2) except it excludes the factor $1 - f$ in the second term. In contrast, most of regression estimator literature presents the expression of the variance for the SRS case ignoring completely the second term in (2) (Fuller, 2009). We will show that this term may play a role later that is not reflected in the regression estimator approximation.

The expression of the variance of poststratification estimates in more complex designs can be derived from (4) using different models; however, an expression for complex designs is not presented explicitly. Furthermore, in regression theory, regression estimators in designs other than SRS may incorporate more information other than the poststratum total and poststratum membership. This information may include information

such as sampling stratum that is not used in the classical poststratification approach. Since we focus on the properties of the poststratification estimator, all regression estimators with information other than the poststrata are out of the scope of the study even if they produce more efficient estimates in large samples.

1.4 Poststratification and ANOVA decomposition

In addition to the analysis of the poststratification estimator as a regression estimator with a group model, Särndal et al. (1992) provide an expression to evaluate the gains of precision of poststratification under simple random sampling using Analysis of Variance (ANOVA) methods. The reduction of variance when an estimate from a SRS design is poststratified is

$$\gamma \approx 1 - R^2, \tag{5}$$

where R is the correlation coefficient for the regression of y_k's on the poststratum indicators g's, and the coefficient of determination R^2 is a measure of the homogeneity of the poststrata corresponding to the proportion of the variance that is explained by the ANOVA model[2]. The reduction of variance of a poststratified estimate depends on how the variability of y_k is explained by the poststrata. Large reduction of variance can be achieved when R^2 is large. Equation 5 ignores the increase of variance due to the fact that the realized sample in the poststrata is random.

1.5 Poststratification and Unequal Probability Sampling

Knottnerus (2003) takes a different approach to poststratification with a special case of unequal probability design. In this design, the inclusion probability π_k proportional to a discrete variable X_k that assumes G mutually different values. This design resembles a stratified sample from a population with G strata where all elements within the same strata have the same inclusion of probability. Furthermore, for $n_g > 0$ each individual stratum is treated as a SRS sample of fixed size n_g. Knottnerus provides the expression of the variance for this case as

$$V\left(\hat{y}_{ps}\right) \approx \sum_g N_g^2 \left(1 - \pi_g + \frac{V\left(n_g\right)}{N_g \pi_g^2}\right) \frac{S_g^2}{N_g \pi_g}. \tag{6}$$

This result is important because it opens the doors for the concept of an equivalent design. Although this concept is described in more detail in Section 3, under certain conditions, this unequal probability design has the same variance as that of a proportionately allocated stratified sample.

[2] The intraclass correlation ρ (Cocharn, 1977) that measures the homogeneity of elements within clusters is related to coefficient of determination R^2 when the poststrata are seen as clusters.

2. Poststratified Estimates of Totals in Stratified Designs

2.1 Poststratification of Totals in Simple Random Sample Designs

One way of expressing the effect of poststratification for the estimator of a total under simple random sampling is as the difference between the approximate variances of the unadjusted estimator and the poststratified estimator. This difference is computed as

$$V\left(\hat{y}^{srs}\right)-V\left(\hat{y}_{ps}^{srs}\right)\approx\frac{N^2}{n}\sum_g W_g\left(\overline{Y}_g-\overline{Y}\right)^2. \qquad (7)$$

Equation 7 assumes that the observed sample in the poststrata is large so the increase of variance due to the fact that the sample size in the strata is random is negligible. This result shows that the difference of the variance is a function of the squared differences on the mean of the poststrata and the total mean. The difference is always positive or zero (when the poststratum means are equal to the overall mean). In other words, the variance of the poststratified estimator is always lower than the variance of the unadjusted estimator if there are differences in poststratum means.

As example using the PUMA population for this study (see Appendix A for a description of the characteristics of the population), we evaluate the difference of the variance of the poststratified estimate of total income when we poststratify to variables in three cases: 1) ethnicity: Hispanic and Non-Hispanic, 2) household tenure categories: own and rent/other, and 3) the combination of ethnicity and household tenure. In this example, the sample size is 3,000 persons. We also compute the ratio of variances of a poststratified total relative to the variance of the total from a SRS design as

$$\gamma=\frac{V\left(\hat{y}_{ps}\right)}{V\left(\hat{y}\right)}=1-\frac{V\left(\hat{y}\right)-V\left(\hat{y}_{ps}\right)}{V\left(\hat{y}\right)}. \qquad (8)$$

Table 1 shows the variance and the ratio of variances, γ. When the sample size is large the reduction in variance of a poststratified estimate of total does not depend on the sample size. In this example, there is no reduction in variance in total income if the estimate is poststratified to ethnicity. There is a 7 percent reduction when the sample is poststratified to household tenure, and the marginal gain of poststratifying to both ethnicity and household tenure is negligible.

Table 1: Difference in Variance and Reduction of Variance of a Poststratified Estimate of Total and the Unadjusted Estimate for a Simple Random Sample Design for the PUMA Population

Poststrata	*Difference of variance (in millions)* $V\left(\hat{y}\right)-V\left(\hat{y}_{ps}\right)$	*Ratio of variance of poststratified total to unadjusted total* (γ)
Ethnicity	59,601,364,323	0.9951
Household tenure	803,110,081,192	0.9341
Ethnicity * household tenure	821,319,039,353	0.9326

2.2 Poststratified Totals in Stratified Simple Random Samples

We begin by examining stratified simple random sampling, where the strata are identified by $h = 1, .., H$. The poststratified estimator is computed as

$$\hat{y}_{ps}^{st} = \sum_g \sum_h \frac{N_h N_g}{n_h \hat{N}_g} \hat{y}_{hg} = \sum_g \sum_h \frac{N_h N_g}{n_h \hat{N}_g} \sum_{k \in s_g} d_k y_k \tag{9}$$

This estimator is a linear function of separate ratio estimators. Using the general expression of variance of the regression estimator with a stratified design, the approximate variance is

$$V\left(\hat{y}_{ps}^{str}\right) \approx \sum_g \sum_h \frac{N_h^2}{n_h} W_{gh} \left\{ S_{gh}^2 + \left(\overline{Y}_{gh} - \overline{Y}_g\right)^2 \right\} \tag{10}$$

where $W_{gh} = N_{gh} N_h^{-1}$ is the proportion of the population in poststratum g in stratum h, and $S_{gh}^2 = \sum_{k \in gh} \left(y_k - \overline{Y}_{gh}\right)\left(N_{gh} - 1\right)^{-1}$ the population variance in the cell for the intersection of poststratum g and stratum h.

The variance of a stratified estimate of a total has two components. The first component is the population variance in the cell of the intersection of the sampling stratum and poststratum. The second component includes the squared difference of the means in the cell for intersection of sampling stratum and poststratum and, the mean of the poststratum within the sampling strata.

Using (10) we can compute an expression of the difference of the variances of the poststratified total from a stratified design and the variance of the unadjusted estimator as

$$Var(\hat{y}^{st}) - Var(\hat{y}_{ps}^{st}) \doteq \sum_g \sum_h \frac{N_h^2}{n_h} W_{gh} \left\{ \left(\overline{Y}_{gh} - \overline{Y}_h\right)^2 - \left(\overline{Y}_{gh} - \overline{Y}_g\right)^2 \right\} \tag{11}$$

In contrast with simple random sampling, with stratified simple random sampling poststratification can either increase or decrease the variance of the estimated total. The effectiveness depends largely on the relationship between the strata and poststrata means.

As example using the PUMA population, we evaluate the difference of the variance of the poststratified estimate of total income for a stratified total with strata created using ethnicity (Hispanic and Non-Hispanic) and poststratifying to household tenure (own and rent/other). The sample size is 2,000 from the non-Hispanic stratum and 1,000 from the Hispanic for a total of 3,000 persons in the sample. The difference in variance is 888,407,324,997 and the ratio of variances is 0.9381. In this example, there is a reduction of variance of the total of 6.2%.

3. Functions, Transformations, and Identities

Although the results presented in the previous sections are informative, they do not provide insights for rules of thumb for evaluating gains (or losses) of variance when

poststratification is used. The main difficulty is finding simple mathematical expressions; and the mathematical complexity increases for estimators other than totals or when other complex designs are studied. In order to overcome this difficulty we introduce two concepts: 1) functional identities and 2) transformations. The objective of these tools is to simplify the complexity of the mathematical operations so we can arrive to simplified expressions or approximations for the variances and reduction of variance when poststratification is used.

3.1 Functions and Transformations of an Estimator

The first tool we introduce is the concept of a transformation. Mathematically, a transformation is a function that establishes a relationship between a given set of elements (i.e., domain) and another set of elements (i.e., range). In survey sampling there are functions that use (transform) the observed data drawn from using a particular design to produce an estimate (i.e., total, mean, proportion, etc.). In other words, estimators such as totals, means, proportions are themselves functions of the observed data[3].

With transformations we need distinguish among elements of the estimate such as the sample design, sample size, etc.; hence we depart from the conventional notation used to describe an estimator. For example, we first define the notation for a total for a SRS sample as

$$\hat{y} = T(\mathbf{y}, srs, n), \tag{12}$$

where T is the function that takes the observed data $\mathbf{y}$ and expands it by the factors N/n and sums it from a SRS design with replacement. In another example, the estimate of the mean for the stratified design in the proposed notation is defined as a function as

$$\hat{\bar{y}}^{str} = M(\mathbf{y}_h, str.H, \mathbf{n}_h) \tag{13}$$

where M is the mean function for a stratified design with H strata and with a vector $\mathbf{n}_h$ with the stratum sample sizes.

Another transformation we introduce is useful for the estimation of means (although it can be generalized to other ratio statistics). For example, the estimate of a domain mean is computed as

$$\hat{\bar{y}}_d = \frac{\hat{y}_d}{\hat{N}_d}. \tag{14}$$

Since this is not a linear estimator but a ratio estimator, in order to compute the variance of this estimate, we use the Taylor series approximation to obtain the linearized approximation of $\hat{\bar{y}}_d$, $e_k = \frac{\delta_k(d)}{\hat{N}_d}\left(y_k - \frac{\hat{y}_d}{\hat{N}_d}\right)$ (Särndal, et al., 1992), where $\delta_k(d) = 1$ if $y_k \in d$, $\delta_k(d) = 0$, otherwise. With the expanded notation, we can rewrite the estimator

[3] A formal framework for the concept of transformation is out of the scope if this paper and will be the topic of a future paper

of the domain mean $\hat{\bar{y}}_d = M(\mathbf{y}_d, srs, n)$, $\mathbf{y}_d$ is the vector of observed data where $\{y_{kd} = \delta_k(d) y_k\}$. It is easy to verify that the approximate variance of $\hat{\bar{y}}_d$ can be written as

$$V(\hat{\bar{y}}_d) = V(M(\mathbf{y}_d, srs, n)) \approx V(T(\mathbf{e}, srs, n)) \tag{15}$$

In other words, the variance of the domain mean is approximated by the variance of the residuals. Notice that the transformation in this example maintains the same design but the form of the estimate and the data used to compute the estimate are not the same as the original estimate. This result is not new (Tepping, 1968). The main motivation for a transformation like this one is that the complexities of the nonlinear estimate are replaced by a linear estimate that is mathematically easier to study.

Although the transformation functions discussed in this section looks like a mere notational convenience, the usability of this notation will be more evident in the following sections.

3.2 Equivalent designs

The second concept we introduce is mathematical functional equivalency. The motivation is that is there are some problems that are difficult to solve or are algebraically un-tractable in their original representations. Through the use of functional identities (or substitutions) we can replace expressions by equivalents that are easier to manipulate and solve. Identities have been used in sampling theory before but without a formal framework. Most of these identities are not generally used in estimation but as tools for evaluation of sample designs. Some examples are the use of concepts such as design effect and effective sample size. The design effect is defined as the ratio of the variance of estimate of the complex design to the variance of a simple random sample design with the same sample size. The effective sample size is defined as the ratio of the nominal sample size of the complex design and the design effect. An implication of these concepts is that there is a SRS design with a sample size with the same value as the effective sample size that produces the same variance as the complex design (Kish, 1995). In other words, this hypothetical SRS sample design and the complex designs are *functional equivalent* with respect to the variance of the estimate.

There are many designs that are equivalent at different levels (i.e., with respect to means, variances, etc.); however, we will focus on a class of equivalent designs centered on a proportionally allocated design. The reason for reducing any design to a proportionally allocated design will become evident in the following section.

Theorem 1. Let $\mathbf{y}$ be the sample vector of observed values drawn using a stratified design p_1 with a sample size $n = \sum_h n_h$. Another sample $\mathbf{y}^*$ is drawn using a stratified proportionally allocated design p_2 from the same strata with a total sample n^* such as

$$V(T(\mathbf{y}, p_1, n)) = V(T(\mathbf{y}^*, p_2, n^*)) \tag{16}$$

These two designs are *equivalent with respect to the variance of* $\mathbf{y}$ for an estimated total.

Proof: The variance of the estimate of total of the fist design is computed as

$$V\left(T\left(\mathbf{y},p_1,n\right)\right)=\sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\frac{N_h^2}{n_h}S_h^2, \tag{17}$$

and the variance of the second design is computed as

$$V\left(T\left(\mathbf{y}^*,p_2,n^*\right)\right)=\left(1-\frac{n^*}{N}\right)\frac{N^2}{n^*}\sum_{h=1}^{H}W_hS_h^2. \tag{18}$$

Substituting expressions (16) and (17) in (15) and solving for n^*, we find that the sample size of the second design that produce the same variance as the first design with a sample n is

$$n^*=\left(\frac{\displaystyle\sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h}+\frac{1}{N}}{N^2\displaystyle\sum_{k=1}^{K}W_hS_h^2}\right)^{-1}. \tag{19}$$

This expression is a generalization of effective sample size. If we compute the design effect as $DEFF=n/n^*$, and we assume $N>>n$, N is large, and the stratum variance is constant, it is easy to verify that

$$DEFF=\left(\sum_{h=1}^{H}W_kr_h\right)\left(\sum_{h=1}^{H}\frac{W_h}{r_h}\right), \tag{20}$$

where $r_h=n_h/N_h$ is the stratum sampling rate. As shown by Theorem 1, Kish's *DEFF* defined in (20) correspond to the ratio of the variance of a stratified design to a proportionally allocated design, and not to a simple random sample design.

3.3 Poststratification of a proportionally allocated design

The next the expression is for the variance for a proportionally allocated stratified design. This expression is just a generalization of the simple random design case and the variance is derived in the same way, conditioning first on the expected sample size in the poststrata and the computing the variance of this expected sample. The vector of the sample size in the poststrata is $\mathbf{n}_G=\left(n_1,...,n_g,...,n_G\right)'$ where n_g is the sample size in the poststratum g is $n_g=\sum_{h=1}^{H}n_{gh}$ where n_{gh} is the sample size in the intersection of sampling stratum h and poststratum g.

Since the sample n is proportionally allocated across the strata and poststrata, it is easy to verify that the expected value n_g is

$$E\left(n_g\right)=n\frac{N_g}{N}=\sum_h n_h\frac{N_{gh}}{N_h}. \tag{21}$$

As in the SRS case, the first component of the expression of the variance of poststratified total from a proportionally design is derived conditioned on the sample $\mathbf{n}_G = \left(n_1,...,n_g,...,n_G\right)'$ as

$$V\left(\hat{y}_{pst}^{str}\right) = E\left(V\left(\hat{y}_{pst}^{str} \mid \mathbf{n}_G\right)\right) + V\left(E\left(\hat{y}_{pst}^{str} \mid \mathbf{n}_G\right)\right) = E\left(\sum_{g=1}^{G} W_g \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_g^2 \mid \mathbf{n}_G\right). \qquad (22)$$

The unconditional variance is computed using an approximation to $E\left(\frac{1}{n_g}\right)$, however, unlike in the simple random case, n_g is the sum of random variates n_{gh} each one with a hypergeometric distribution with parameters $\left(n_h, N_{gh}, N_h\right)$. Using the Taylor's series approximation, we have

$$E\left(\frac{1}{n_g}\right) = E\left(\frac{1}{\sum_{h=1}^{H} n_{gh}}\right) \approx \frac{1 + V\left(\sum_{h=1}^{H} n_{gh}\right)}{E\left(\sum_{h=1}^{H} n_{gh}\right)}. \qquad (23)$$

Since the samples are selected independently within sampling stratum, $V\left(\sum_{h=1}^{H} n_{gh}\right) = \sum_{h=1}^{H} V\left(n_{gh}\right)$. The expression of the variance is

$$V\left(\hat{y}_{pst}^{str}\right) \approx \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \sum_{g=1}^{G} W_g S_g^2 + \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{n_h^2} \sum_{g=1}^{G} \left(1 - \frac{N_{gh}}{N_h}\right) S_{gh}^2 \qquad (24)$$

As in (2), the variance has two components. The first component is the expression of the variance as if the sample had been drawn proportionally by poststratum. Although the second component is more complex, it still reflects the increase of variance due to the variability of the sample in the cell formed by the intersection of the sampling stratum and poststratum.

Using the expanded notation, we can rewrite (24) as the following identity that links the variances of proportionally allocated stratified design with and without poststratification:

$$V\left(P\left(T\left(\mathbf{y}, prop.str.H, n\right), G\right)\right) = V\left(T\left(\mathbf{y}, prop.str.G, n\right)\right) + \Delta \qquad (25)$$

where $\Delta = \left(1 - \frac{n}{N}\right) \frac{N^2}{n^2} \sum_{h=1}^{H} \sum_{g=1}^{G} \left(1 - \frac{N_{gh}}{N_h}\right) S_{gh}^2$. We will use these results in the following section.

4. Gains of poststratification in stratified designs

4.1 Poststratified totals from a stratified design

In this section we compute the gains of poststratification for totals when the sample is drawn from a stratified design. In particular, we are interested in the ratio of variances as

defined in (8) but with the variance of the estimated total from the stratified design in the denominator. Using the previous results the ratio γ of variances becomes

$$\gamma = \frac{V\left(P\left(T\left(\mathbf{y}, str.H, n\right), G\right)\right)}{V\left(T\left(\mathbf{y}, str.H, n\right)\right)} = \frac{V\left(P\left(T\left(\mathbf{y}, prop.str, H, n^*\right), G\right)\right)}{V\left(T\left(\mathbf{y}, prop.str, H, n^*\right)\right)} = \frac{V\left(T\left(\mathbf{y}, prop.str, G, n^*\right)\right) + \Delta}{V\left(T\left(\mathbf{y}, prop.str, H, n^*\right)\right)}.$$

As in Särndal et al. (1992), we can express the variances in terms of the coefficient of determination R^2. We modify the notation of R^2 to indicate if the coefficient is computed using the strata or poststrata. Since

$$\frac{V\left(T\left(\mathbf{y}, str, H, n^*\right)\right)}{V\left(T\left(\mathbf{y}, srs, n^*\right)\right)} \approx 1 - R^2\left(\mathbf{y}, H\right) \text{ and } \frac{V\left(T\left(\mathbf{y}, str, G, n^*\right)\right)}{V\left(T\left(\mathbf{y}, srs, n^*\right)\right)} \approx 1 - R^2\left(\mathbf{y}, G\right),$$

the ratio γ reduces to

$$\gamma = \frac{1 - R^2\left(\mathbf{y}, G\right) + \Delta / V\left(T\left(\mathbf{y}, srs, n^*\right)\right)}{1 - R^2\left(\mathbf{y}, H\right)} \approx \frac{1 - R^2\left(\mathbf{y}, G\right)}{1 - R^2\left(\mathbf{y}, H\right)}, \tag{26}$$

when n^* is large. Equation (26) shows that the reduction of variance when a stratified estimate is poststratified depends on how well the sampling strata or poststrata explain the variability of $\mathbf{y}$. For example, if the sampling strata produce a larger R^2 than the poststrata, then poststratifying the estimate increases the variance.

As a numerical example using the PUMA data, Table 2 shows the gains in poststratification for the estimate of total income when the sample of 3,000 persons is drawn disproportionally from a frame stratified by ethnicity. In the table, the first column shows the relative sampling rate of the Hispanic stratum with respect to the non-Hispanic stratum. The second column shows the gains in poststratification, γ, as computed in (26) when the estimates of total income are poststratified to totals by household tenure (own, rent/other). The third column shows the estimated value of γ using repeated sampling (i.e., simulation) with 5,000 runs. These results show that the gains achieved using poststratification do not depend on the sample allocation as long the effective sample size in the cell for the sampling stratum/poststratum is large. However, poststratification will not recover the gains of proportional allocation sample as in the SRS case if the sample is drawn from a stratified design.

Table 2: Gains from Poststratification to Household Tenure of Estimates of Total Income in the PUMA Population Stratified by Ethnicity

Relative sampling rate	γ	*Simulation*
4.00	0.9387	0.9366
1.00	0.9387	0.9418
0.25	0.9387	0.9454

We now take a look at the increase of variance in poststratified totals from stratified designs $V\left(P\left(T\left(\mathbf{y}, str.H, n\right), G\right)\right) = V\left(T\left(\mathbf{y}, prop.str, G, n^*\right)\right) + \Delta$ due to, Δ, that is generally ignored. In the stratified case we can compute this term. If we assume a constant

population variance across strata and poststrata and negligible finite population factors, the increase of variance can be expressed as

$$\Delta' = \frac{GH - 1}{GH\ \bar{n}^*_{gh}}, \tag{27}$$

where $\bar{n}^*_{gh} = n^* / (GH)$ is the average number of effective sampled units per stratum/poststratum. This expression is a generalization of the expression for poststratification in the SRS case (i.e., one stratum) given in Cochran (1970), with for the inclusion of the number of poststrata and the effective sample. The conditions for large increases of variance are also generalizations of those in Cochran. However, it is often overlooked in practice that the variability depends on the effective sample size of the cell for the intersection of sampling stratum/poststratum. Practitioners often examine the nominal sample in the poststratum as in the SRS case. In the stratified case, this practice may not be a problem because this term is only important when $\bar{n}^*_{gh}$ is less than one, which is very rare in stratified designs. However, this situation may be true in other design such as cluster sampling.

4.2 Poststratification of means and domains in a stratified design

In the stratified case, the estimate of the mean $\bar{Y}$ is the estimate of total Y divided by N. Therefore the results of the gain/losses due to poststratification for stratified totals apply to stratified means. Since the $1/N^2$ appears in the numerator and denominator in the expression for γ, the gains or losses due to poststratification in stratified means are the same as those achieved in totals. This is also true for estimates of domain totals.

4.3 Poststratification of domain means in a stratified design

A more interesting case is the case of poststratified domain means from stratified samples. Using previous results, we can compute the reduction of variance of the domain mean from a SRS design as

$$\gamma = \frac{V\left(P\left(M\left(\mathbf{y}_d, srs, n\right), G\right)\right)}{V\left(M\left(\mathbf{y}_d, srs, n\right)\right)} \approx \frac{V\left(P\left(T\left(\mathbf{e}, srs, n\right), G\right)\right)}{V\left(T\left(\mathbf{e}, srs, n\right)\right)} = \frac{V\left(T\left(\mathbf{e}, prop.str, G, n\right)\right) + \Delta}{V\left(T\left(\mathbf{e}, srs, n\right)\right)}$$

If the sample n is large, then the ratio of variances is $\gamma \approx 1 - R^2\left(\mathbf{e}, G\right)$. This result is similar to (5) (also see Särndal et al., 1992) but in terms of the residuals. This result shows that there is always a reduction of variance in a poststratified domain mean from a SRS design; the reduction of variance due to poststratification depends on how well the variability of the residuals is explained by the poststrata.

The results can be generalized to compute the gain or loss due to poststratification for stratified domain estimates. The ratio of variances is

$$\gamma = \frac{1 - R^2\left(\mathbf{e}, G\right)}{1 - R^2\left(\mathbf{e}, H\right)} \tag{28}$$

This result is similar to (5) but the in terms of residuals instead of the variable y.

5. Conclusions and Further Research

Simple rules of thumb to evaluate the gains due to poststratification for estimates from complex designs are difficult to derive due to the mathematical complexity. However, with the introduction of simple concepts we generalize results already described in the literature to the stratified simple random design. These results show that poststratification does not always reduce the variances, and the gains or losses of efficiency depends on how well the sampling strata and poststrata explain the variability of the variable of interest. These results also show that for gain (or losses) due to poststratification for domains totals and domains means are not necessarily the same.

The current rules of thumb used by sampling practitioners when poststratification is implemented may not applicable in estimates from stratified designs. Although this will be examined more detail in a future study, the conditions where the increase of variance due to the second term are not very likely to occur in practice for most stratified designs. This may not hold for other complex designs.

References

Cochran, W.G. 1977. Sampling Techniques, 3rd edition. John Wiley & Sons, Inc: New York.

Fuller, W.A. 1966. Estimation employing poststrata. *Journal of the American Statistical Society*, 61, 1172-1183.

Fuller, W.A. 2009. *Sampling Statistics*, John Wiley & Sons: New York.

Hartley, H.O. 1962. Multiple Frame Surveys. In *JSM Proceedings,* Social Statistics Section. Alexandria, VA: American Statistical Association. 203–206.

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

Knottnerus, P. 2003. Sample Survey Theory: Some Pythagorean Perspectives. Springer-Verlag: New York.

Kott, P.S. 2006, Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology,* 32(2), 133-142.

Särndal C.-E., Swensson B. and Wretman J.H. 1992. Model Assisted Survey Sampling. New York: Springer.

Tepping, B. 1968. The Estimation of Variance in Complex Surveys. In *JSM Proceedings,* Social Statistics Section. Alexandria, VA: American Statistical Association, 11-18.

Thompson, S. K. 1992. Sampling. John Wiley & Sons: New York.

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

Research Triangle Institute 2008. *SUDAAN Language Manual, Release 10.0* Research Triangle Park, NC: Research Triangle Institute.

Appendix

The source of the data is the 5 percent Public Use Mircodata Sample (PUMS) from the 2000 U.S. Decennial Census for the West and South. The PUMS data consist of information located geographinc areas designated *Public Use Microdata Area* (PUMA) code. Table A shows the characteristics of this population for the variables used in this study.

Table A: Distribution of Income Population

Categorical variable	*Number of persons*	*Mean*	*Standard deviation*
All ethnicity	5,360,596	52,436	35,669
Hispanic	776,738	46,376	31,859
Non-Hispanic	4,583,858	53,463	36,175
Household tenure			
Own	3,751,226	58,434	36,561
Other	1,609,370	38,456	29,032