VARIANCE ESTIMATION FROM CALIBRATED SAMPLES

Douglas Willson, Paul Kirnos, Jim Gallagher, Anka Wagner

National Analysts Inc.
1835 Market Street, Philadelphia, PA, 19103

Key Words: Calibration; Raking; Variance Estimation.

1 Introduction

Survey researchers often adjust preliminary survey analysis weights so that sample estimates match known control totals for auxiliary variables. These adjustments, called *raking* or *calibration*, are attractive in that the resulting statistical estimates have desirable properties, including reduced bias and increased efficiency in some circumstances. Adjusting survey weights to match external control totals also confers benefits in terms of consistency, which may be important in situations where the survey belongs to a larger group of inter-related surveys, or when alignment across estimates from different sources is required. Over the years, many different approaches for raking or calibration have been proposed. Singh and Mohl (1996) provide a detailed description for many of these methods.

Appropriate estimates of variances for statistics from calibrated samples can be computed using a variety of different methods. Deville and Sarndal (1992) show that variances of many common calibration estimators can be estimated using standard Taylor series formulae for the generalized regression estimator. Replication procedures such as the jackknife and the bootstrap can also be employed.

Unfortunately, standard commercially-available software that use Taylor-series methods (such as SUDAAN, SAS (Proc Surveymeans),or STATA) typically do not provide the appropriate estimates for calibrated samples. For the replication-based estimates, the situation is somewhat more complicated. One advantage that is often emphasized with replication-based approaches is that post-survey weighting adjustments such as calibration/raking and post-stratification can be included in construction of the replicate weights, thereby providing a "true" estimate of the variability of estimates in repeated samples. From the perspective of calibration, each replicate can theoretically be calibrated to the control totals. However, with the exception of WESVAR, most standard software does not provide the option to calibrate replicate subsamples. WESVAR allows for calibration during replication using the method of iterative proportional fitting. Other methods, including those involving range restrictions (see Singh and Mohl (1996) or Section 2) are not supported. As a result, calibrating replicate weights to properly reflect the chosen raking method may require special purpose software.

For secondary analysis using survey data from calibrated samples, it is also often the case that detailed information concerning the sample and raking targets are not provided, yet this information is required to calculate the Taylor series variance estimates properly. For replication-based procedures, it may be difficult to properly calculate variance estimates from calibrated samples unless recalibrated replicate weights are provided with the survey dataset. When available variance estimation procedures do not properly reflect the calibration, variance estimates may be biased, and inferences may be altered because the calibration information has been ignored.

This paper investigates the magnitude and direction of possible biases in variance estimates when calibration information has been ignored. In particular, we compare traditional Taylor series and replication-based estimates with several convenient approximations that can be computed using standard software with no information concerning external calibration totals. The comparisons are made using simulated samples drawn from a hypothetical population. The paper will proceed as follows: Section 2 provides some background concerning raking methods. Section 3 describes several methodologies for variance estimation with calibrated samples, as well as some convenient approximations that do not directly incorporate information concerning the calibration constraints. Section 4 constructs a hypothetical population and conducts a simulation study investigating the prop-

erties of the different variance estimates in repeated samples. Section 5 concludes and suggests avenues for future research.

2 Background

Deville and Sarndal (1992) and Deville, Sarndal and Sautory (1993) (henceforth DSS) consider the following notation. Let n, N denote the sample size and population size respectively. Let d_k represent the usual design-based survey weight (the base weight) for respondent k. Let y_k be the value of a variable of interest for the k^{th} population element, and let x_k=$\{x_{k1}, ...x_{kJ}\}$ be a vector of J auxiliary variables. For the auxiliary variables, we assume that the population totals or benchmark constraints are known, *i.e.* τ_j=$\sum_{i=1}^{N} x_{ij}$.

The basic idea behind calibration is to develop new weights $\{w_k$, k=1...n$\}$ for each respondent such that the survey sample produces estimates that match the population or benchmark totals. Following D-S, this can be operationalized as a minimum-distance problem, with different calibration estimators employing different distance measures.

To illustrate, DSS consider distance measures $G_k(w,d)$ satisfying certain regularity conditions with $g_k(w,d)$= $\partial G_k/\partial w$. Calibration estimators are chosen to minimize distance measured as $\sum_{k=1}^{n} G_k(w_k, d_k)$ subject to the J calibration constraints. Let λ be a vector of lagrange multipliers. It follows that

$$g_k(w_k, d_k) - x_k'\lambda = 0. \tag{1}$$

In what follows, it is useful to write this as

$$w_k = d_k F_k(x_k'\lambda). \tag{2}$$

where F=G^{-1}.

It is informative to examine the minimization using G(w,d)=$\sum_{k=1}^{n}(w_k$-$d_k)^2/d_k$ which corresponds to the linear regression or unrestricted modified minimum chi-square calibration method. In this situation, Equation (2) implies that

$$w_k = d_k(1 + x_k'\lambda) \tag{3}$$

where

$$\lambda = (\sum_{k=1}^{n} d_k x_k x_k')^{-1}(t_x - \hat{t_{x\pi}}) \tag{4}$$

where t_x is the estimator of population total for x using the calibrated weights, and $\hat{t}_{x\pi}$ is the usual design-based estimator. The generalized regression estimator of the population total for a variable y can be written as

$$\hat{t}_{yreg} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B} \tag{5}$$

where

$$\hat{B} = (\sum_{k=1}^{n} d_k(x_k x_k')^{-1} \sum_{k=1}^{n} d_k x_k y_k \tag{6}$$

The variance of the generalized regression estimator is

$$\sum_k \sum_l (\pi_{kl} - \pi_k\pi_l)\pi_{kl}^{-1}(e_k d_k)(e_l d_l) \tag{7}$$

where $e_k = y_k - x_k'B$. DSS show that estimators from a broad family of distance function are asymtotically equivalent to the generalized regression estimator, and have this variance.

3 Variance Estimators

3.1 Taylor Series

Estimates of the variances in multistage designs are typically computed assuming that first-stage sampling units are selected with replacement. For the work that follows we will only consider simple stratified designs, for which an estimate of the variance $\hat{V}(\hat{Y}_{TS})$can be constructed as

$$\sum_{h=1}^{H} \frac{n_h}{n_h - 1}(sum_{k=1}^{n_h}(\sum_{k\in h} d_{hk}e_{hk} - \frac{1}{n_h} \sum_{k=1}^{n_h} \sum_{k\in h} d_{hk}e_{hk})^2 \tag{8}$$

for strata $h = 1...H$ with n_h representing the sample size in stratum h. Note also (as in Stukel, Hidiroglou, and Sarndal (1996)), this estimator is not the true Taylor series estimator because of the assumption of with-replacement sampling, but we will refer to it as the Taylor series estimator for historical reasons.

While this estimator uses the design-based weights, an improvement can be made by substituting the weights from the calibration estimator w_{hk} for d_{hk} that will depend on the distance function used in the raking algorithm. For the empirical work presented in Section 4, we consider two different raking algorithms: 1) Unrestricted Modified Discriminant Information (MDI-u), also called "raking ratio" or "iterative proportional fitting"; and 2) Restricted Modified Discriminant Information (Method 6 in Singh and Mohl). MDI-u is perhaps the most widely used calibration method. It is attractive because the calibrated weights are non-negative, improving on the

regression estimator where negative weights are possible. The approach is also guaranteed to converge as the number of iterations increases, which makes it attractive for resampling-based approaches to variance estimation. The second uses the same measure of distance but imposes range restrictions on the degree of relative movement between the original and final weigths. Range restrictions are motivated by the observation that MDI-u often produces large weights which may dominate some analyses, particularly when domains are considered.

For MDI-u, the total distance function can be written as

$$G_{MDI-u}(w,d) = \sum_{k=1}^{n}(w_k log(w_k/d_k) - w_k + d_k) \quad (9)$$

For MDI-r, range restrictions are represented as $L<w_k/d_k<U$ for lower bound U and upper bound L, where L<1<U, and the observation-specific distance function can be written as

$$G_{MDI-r}(w_k,d_k) = G_{MDI-u}(w_k,d_k) \quad (10)$$

for $L < w_k/d_k < U$ and

$$G_{MDI-r}(w_k,d_k) = \infty \quad (11)$$

otherwise. In our empirical work, the two variance estimates corresponding to these specific distance functions will be denoted $\hat{V}_{TS}(\hat{Y}^u)$ and $\hat{V}_{TS}(\hat{Y}^r)$.

3.2 Jackknife

The basic idea behind the jackknife is to drop one or more observations from the sample and to recalculate the estimates from the remaining observations. This process is repeated until all observations have been dropped. If $\hat{\theta}$ is the survey estimate from the entire sample, and $\hat{\theta_{-i}}$ is the estimate for the sample with observation i removed, the jackknife estimate of the variance is calculated as

$$\hat{V}_J(\hat{\theta}) = \frac{1}{n-1}\sum_{\forall i}(\hat{\theta}_{-i} - \hat{\theta})^2 \quad (12)$$

In stratified samples, it is important to reflect the stratification in the replications, and to calculate the variances within each strata. (Wolter, 1985) Rust (1985) suggests the following formula for h=1...H strata:

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^{H}\frac{n^h - 1}{n^h}\sum_{\forall i}(\hat{\theta}^h_{-i} - \hat{\theta}^h)^2 \quad (13)$$

In our empirical work, we denote jackknife estimates from calibrated samples using MDI-u as $\hat{V}_J(\hat{Y}^u)$ and using MDI-r as $\hat{V}_J(\hat{Y}^r)$. Both of these estimates are recalibrated for each replicate.

3.3 Some Convenient Alternatives

In this section we present some variance estimators that use the calibrated weights but do not directly employ the calibration constraints in the analyses. The first alternative uses Taylor series formulae available in most standard software, but ignores the calibration information entirely. We assume that stratum identifiers are available, as well as the calibrated weights. In this situation, the variance of the total is calculated as:

$$\hat{V}^a_{TS}(\hat{Y}_T) = \sum_{h=1}^{H}(1-f_h)n_h S_h^2 \quad (14)$$

where S_h^2 is the variance of the appropriate linearized value, *i.e.*

$$S_h^2 = \frac{1}{n_h - 1}\sum_{k\in h}(w_k y_k - \bar{y_k}^h)^2 \quad (15)$$

For the empirical work we present below, Taylor series estimates that ignore the calibration information but use corresponding calibrated weights are denoted $\hat{V}^a_{TS}(\hat{Y}^u)$ and $\hat{V}^a_{TS}(\hat{Y}^r)$. For the jackknife, we also consider the convenient approximations $\hat{V}^a_J(\hat{Y}^u)$ and $\hat{V}^a_J(\hat{Y}^r)$, where the replicate samples are not recalibrated in each iteration. (The weights are adjusted to account for the dropped observation in each replicate however).

4 Simulation Study

4.1 Design

We investigate the properties of the different variance estimators in repeated samples using a hypothetical population. We do not attempt to provide a comprehensive investigation of the behavior of alternative variance estimators, as in Stukel, Hidiroglou and Sarndal (1996). Instead, we focus on the relative performance of the convenient approximations when compared with the estimates that properly account for the calibration.

A hypothetical population was constructed using 20,000 households, sampled without replacement, from the March 2001 Current Population Survey public use dataset. The hypothetical population was stratified into four geographic regions (Northeast, Midwest, South, and West) and three income groups based on household income (<35K,35K-70K,70K+). One thousand simple stratified random samples were selected from the population (each without replacement). Each sample comprised 1008 households, with

equal allocations across the 12 geographic/income strata. For external calibration controls, we considered control totals for total household income and the total number of households with at least one person uninsured. We examined estimates for the total number of children under the age of 18, the total number of married families, and the total of wages and salaries income.

The control information is probably not terrifically useful in explaining (or forecasting) the number of children under 18 or the number of married families. For these variables, one would expect only modest differences between the approximation methods and the properly calculated Taylor series and Jackknife estimates. However, since wages and salaries are a substantial portion of household income (the population correlation is 0.72), there should be significant differences between the approximations and the appropriate variance estimates for this variable.

We report the following statistics that summarize the behavior of the total estimates, and the associated variances estimates:

1) Percent Relative Bias of Total Estimator, $\hat{Y}_T^u$ and $\hat{Y}_T^r$ when compared to the true population value. This is calculated as

$$(\frac{E(\hat{Y}_T) - Y_T}{Y_T}) * 100 \tag{16}$$

where

$$E(\hat{Y}_T) = (\frac{1}{R}) \sum \hat{Y}_T \tag{17}$$

$\hat{Y}_T$=$\hat{Y}_T^u$ or $\hat{Y}_T^r$, and where the average is evaluated over R=1000 samples.

2) Percent Relative Bias of the Variance Estimator, when compared to the true variance. This is calculated as

$$(E(\hat{V}(\hat{Y_T})) - V_{True})/V_{True} * 100 \tag{18}$$

where

$$E(\hat{V}(\hat{Y})) = \frac{1}{R} \sum_{r=1}^{R} \hat{V}_r(\hat{Y}_T) \tag{19}$$

and

$$V_{True} = \frac{1}{R} \sum_{1}^{R} (\hat{Y}_r - E(\hat{Y}))^2 \tag{20}$$

where $\hat{V}_r(\hat{Y}_T)$ is the variance estimate for each subsample for each method, and V_{True} is the true sampling variability of the calibrated estimates as measured by their variability across the 1000 samples.

4.2 Results

Table 1 presents results for the relative bias of the estimates of totals for the two raking methods. Note that the relative bias is extremely small, on the order of one tenth to two tenths of one percent, for both raking methods. It is clear that the neither raking method introduces extreme biases in the total estimates themselves.

Table 1: % Relative Bias, $\hat{Y}_T^u$, $\hat{Y}_T^r$

Total	$\hat{Y}_T^u$	$\hat{Y}_T^r$
Children U18	-0.14	-0.11
Married Families	0.21	0.29
Wages + Salaries	-0.12	-0.11

Tables 2 and 3 present the relative biases for the variance estimates for each of the four methods, using the raked weights from MDI-u and MDI-r. Previous research (Stuckel *et. al*, (1996)) has focused on the differences between $\hat{V}_{TS}$ and $\hat{V}_J$, concluding that the bias associated with the Taylor series estimates is usually larger than the bias for the Jackknife. Our estimates are roughly consistent with this finding, although we note that for both methods there appears to be more bias for the wages and salaries estimate (15.75 percent for MDI-u, 18.66 percent for MDI-r) and the number of children under 18 (9.78 percent for MDI-u, 9.69 percent for MDI-r) than for the number of married families. Wages and salaries is a continuous variables, while the number of children under 18 is a count variable concentrated on a relatively small number of integers. The married family variable is binary (0,1) indicator. The degree of adjustment to the individual weights under raking will be more sensitive to those variables whose values vary across all individuals in each sample. Put differently, matching exactly to a control target when the auxiliary information is continuous may introduce a higher potential for bias than if the auxiliary information is categorical. The average sizes of the biases for $\hat{V}_{TS}$ and $\hat{V}_J$ are consistent with other empirical results (Stuckel *et al.* (1996)) for the categorical variable (married family) but not for the other variables.

The bias for the convenient approximation variance estimates are larger than the biases for the proper Jackknife and Taylor series estimates in most cases, but for children under 18 and married families the biases are of similar order of magnitude. For the wages and salaries variance estimates, the convenient approximations have enormous biases, over 150 percent in each case. The intuition behind this result is clearly illustrated by equation (8), which uses the "raking regression" residuals rather than the values of the variable of interest. For situations where a

significant portion of the variability of the estimate can be "explained" by variations in the control totals, the residuals will exhibit substantially less variability, and the corresponding variance estimates under calibration should be much lower. The convenient approximations miss this correction, and substantially overestimate the variance of the wages and salaries total estimate as a result.

Table 2: % Relative Bias, $\hat{V}(\hat{Y}_T)$, MDI-u

Total	$\hat{V}_{TS}$	$\hat{V}_J$	$\hat{V}^a_{TS}$	$\hat{V}^a_J$
Children U18	14.98	9.78	15.36	4.95
Married Families	7.59	-0.84	9.48	-1.05
Wages + Salaries	-16.89	15.75	164.40	174.22

Table 3: % Relative Bias, $\hat{V}(\hat{Y}_T)$, MDI-r

Total	$\hat{V}_{TS}$	$\hat{V}_J$	$\hat{V}^a_{TS}$	$\hat{V}^a_J$
Children U18	14.01	9.69	15.04	4.63
Married Families	7.31	-0.91	9.43	-1.11
Wages + Salaries	-15.42	18.66	156.31	169.13

5 Conclusions

We have presented results from a small simulation study examining the behavior of variance estimates in calibrated samples. Jackknife and Taylor series variance estimates that properly account for the calibration information were compared with convenient approximations using some of the sample design information but ignoring the calibration totals. For situations where the estimates of interest are primarily unrelated to the calibration information, the convenient approximations using Taylor series or Jackknife methods produced biases that were similar in magnitude to the more complicated procedures that correctly account for the calibration. For situations where the estimates of interest are related to the calibrating variables, the approximations seriously overestimated the true sampling variability of the estimates.

From a practical perspective, the choice between the approximations and the appropriate (but more cumbersome and perhaps time consuming) Taylor series and Jackknife procedures will depend on the relationship between the specific variables under investigation and the control information. In many calibration applications, the control totals may not "explain" a substantial portion of the survey items of interest. For example, surveys that are part of larger information system may be calibrated to ensure consistency across system components. In this situation, calibration will probably have a limited impact on the variances of most items, and the approximations will be acceptable. When the auxiliary information that is used for calibration is highly correlated with items of interest, the fact that the convenient approximations will overstate true sampling variability implies that inferences using these variance estimates will be conservative. If the nature of the calibration controls is known, some analysts may also decide to use the approximations for items where the risk of bias is low, and invest time and resources in the correct calculations when the risk is high.

6 References

Deville, J. C., and Sarndal, C. E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, pp. 376-382.

Deville, J. C., Sarndal, C. E., and Sautory, O. (1993) Generalized raking procedures in survey sampling, *Journal of the American Statistical Association* 88, pp. 1013-1020.

Singh, A.C., and C.A. Mohl (1996) Understanding calibration estimators in survey sampling, *Survey Methodology 22*, pp. 107-115.

Stuckel, D., Hidoroglou, M. A., and Sarndal C. E. (1996) Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization, *Survey Methodology 22*, 117-125.

Wolter, K. M., (1985) *Introduction to Variance Estimation*, New York: Springer.