EMPIRICAL INVESTIGATION OF VARIANCE ESTIMATORS BASED ON VARIOUS DESIGN OPTIONS WHEN A RAKING METHOD IS IMPLEMENTED

Don Jang, Joseph K. Garrett, Mathematica Policy Research, Inc.
Frank W. Piotrowski, William B. Owens, ACNielsen

Key Words: Poststratification, Raking, Variance estimation

I. Introduction

To represent the target population from the final responding samples, weighting adjustments often are necessary. Poststratification is a technique for adjusting survey data by using external data from the census or other large surveys, and it requires a population count for each cell. However, while the marginal distributions for auxiliary variables may be available, the population totals of the complete cross-classifications of all auxiliary variables are often not available. Even with the complete cross-classification counts, a ratio-adjusted poststratification procedure may not be appropriate when some of the classes are so small that they may cause unnecessary weight variation. In such cases, one can collapse those small cells and do poststratification adjustment. However, as the number of domains increases, the number of small cross-classification cells becomes large. This may result in nontrivial bias in estimating marginal totals if those small cells are collapsed. A raking ratio adjustment has been used when only marginal totals are available or too many small cross-classification cells exist even when all cell counts are available.

It is tedious to incorporate the variation due to the raking adjustment in variance estimation. In fact, it is not always feasible to obtain closed forms of variance estimates under the raking procedure. While some research articles (for example, Deville and Särndal 1992) might relate to deriving the variance estimate formula in the context of raking adjustments, it is still difficult to implement the existing procedure on large-scale surveys due to the lack of available software. Many practitioners, however, prefer to use existing software to calculate variance estimates for the survey estimates. This motivated us to approximate variance estimates by either ignoring the raking adjustments or assuming a simplified poststratification adjustment. Variance estimates under postratification can be calculated using standard variance estimation software such as SUDAAN.

In this paper, we apply the idea to ACNielsen's panel survey, a survey for which a complicated raking procedure was implemented. In Section II we briefly introduce poststratification and raking procedures. In Section III we present an illustration using data from ACNielsen's household panel survey.

II. Poststratification and Raking Adjustments

Survey statisticians tend to adjust the design-based weights of respondents so that estimates of certain population totals conform to known values for these totals from external data sources. The primary purpose of such adjustments is to reduce the bias due to incomplete coverage of the target population. A well-known and frequently used method of making this adjustment is post-stratification. In some circumstances, population weighting adjustments may also reduce the variances of the estimates. Poststratification is widely used in household surveys to control the weighted sample totals to known population totals for certain demographic subgroups.

The population totals of the complete crossclassification of the auxiliary variable may not be known, while the marginal distributions for each variable are available. Even if the full crossclassfication is known, the number of respondents in each cell may be small or zero and this can lead to inconsistent and highly variable estimates.

Raking methods are often used to approximate population sizes for some demographic or geographic domains when only marginal totals are readily available or sizes of cross-classification cells are small. Raking procedures often require several iterations to balance the multi-dimension marginal totals. In general, weights are ratio adjusted to conform to the marginal distribution of the first auxiliary variable. These adjusted weights are then ratio adjusted to conform to the marginal distribution of the second auxiliary variable. Continuing in this manner, the first iteration concludes when the last auxiliary variable is fitted. Subsequent iterations need to be done until the weights conform to the marginal distributions of all the auxiliary variables. This approach can provide final weights approximating unknown population counts for cross-classification cells. A raking adjustment can be done using an iterative proportional fitting algorithm. Under general conditions, the algorithm converges to a solution. Further technical details about raking methods are in Kalton and Maligalig, 1991, Oh and Scheuren, 1983, and references cited therein.

Poststratification and raking are two adjustment methods that fit within the broader framework of generalized raking or calibration estimation (Deville and Särndal 1992). The general approach is to adjust the weights so that they satisfy the condition that their sum is equal to the population total for each of the auxiliary variables and that the distance between the unadjusted and adjusted weights is minimized. The condition on the sum of the weights is called a calibration equation. However, this procedure needs auxiliary information like initial sampling weights as well as the final weight. Often the secondary users may not have an access to all the information to implement the analytical variance formula when the raking adjustment is used.

Due to complexity of the variance estimation from the raking adjustments, we consider using relatively simple adjustment options by regarding the final weights as the sampling weights in a sense that no adjustments or poststratification adjustments are assumed. In practice, many survey researchers do this partly because it is easy to implement by using current survey specific software such as SUDAAN. The variance estimators can then be easily calculated using a customary variance estimation method such as Taylor series linearization method if the adjustment option can be simplified.

III. Application to ACNielsen's Panel Survey Data

Our application is to use ACNielsen's household panel survey data. This panel survey is fielded to understand purchasing behavior of consumers for retail store-based transactions. For this panel survey, ACNielsen has implemented a sophisticated raking procedure to account for the demographic and geographic characteristics of the population. To assess the reliability of estimates from the sample properly, it is desirable to have good variance estimates. However, it is complicated to compute sampling variability estimates directly from this panel. For the design aspects, an option for variance estimation is to set reasonable design assumptions that closely approximate sample structure.

1. Sample design

The target population for this panel survey is all households in the contiguous U.S. ACNielsen partitions the population into 20 geographic strata: 16 major market areas and 4 remaining Census regions. They have recruited a panel of 40,000 households matching a selected group of demographic characteristics for each geographic stratum. Of the 40,000 households, 32,425 households met the reporting and editing requirements for second quarter, 1998. We obtain the projected total number of households in this period by summing the projection factors over 32,425 households. This sum is 101,041,273 households. The 16 major markets contain 70.1% of the sample households and account for 33.9% of the total U.S. household population. The remaining four U.S. Census areas contain 29.9% of the sample households, but account for 66.1% of the total U.S. household population. ACNielsen designs the sample to provide local market reports, and thus oversamples in 13 major markets. Conversely, they substantially undersample all four remaining census regions and three major markets.

2. Projection factor construction

To represent the target population from the final responding samples, weighting is necessary. ACNielsen employs an iterative proportional fitting, or raking, algorithm to construct final household weights while approximating unknown population counts for all cross-classification cells.

The raking procedure for this survey requires several iterations to balance the multi-dimension marginal totals. ACNielsen conducts several iterations until the weights conform to the marginal distributions of all the auxiliary variables.

To avoid extreme weights during this process, ACNielsen uses constraints so that weights should not be larger than 4.5 times the average projection factor. Because all auxiliary information, such as demographic data and household size, is available within each geographic area, ACNielsen uses raking procedures independently across the twenty geographic areas.

Specifically, ACNielsen uses a raking technique that incorporates individual and household population counts classified by demographic and geographic characteristics such as county size, household size, female head age, income level, etc. This is done within each major market or remaining U.S. area. Therefore, the resultant weights, called projection factors, can give correct population totals for both market and total U.S. levels. In other words, they project sample households in major markets to total number of households in the corresponding major markets, and they project remaining Census regions' sample households remaining Census regions' universe. Demographic variables in the ACNielsen projection system include:

- Household size - 4 levels
- Household income - 4 levels
- Age of Female Head - 4 levels
- Household Race - 3 levels
- Male Head Education - 4 levels

- Female Head Education - 4 levels
- Head of Household Occupation - 3 levels
- ACNielsen County Size - 4 levels
- Hispanic - Y/N

3. Variance estimation methodology

Two general statistics are considered in our work: totals (total purchases) and ratios (share, buying rate, and penetration rate).

Total. Let $\hat{Y}$ denote an estimator of the population total Y. In the stratified design option we are using,

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h \qquad (1)$$

where

$$\hat{Y}_h = \sum w_{hi} y_{hi} \qquad (2)$$

is an estimator for the h-th post-stratum quarterly total of Y_h, the population total for post-stratum h, w_{hi} is the projection factor of the i-th panel of the h-th post-stratum obtained from the ACNielsen's projection process and y_{hi} is the total purchasing quantity during the quarter. Then, the customary variance estimator of $\hat{Y}$ is the sum of the post-stratum variance estimators, or

$$v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h) \qquad (3)$$

where

$$v(\hat{Y}_h) = n_h^{-1}(n_h - 1)^{-1} \sum_{i=1}^{n_h} (n_h w_{hi} y_{hi} - \hat{Y}_h)^2$$

is a post-stratum variance estimator, and n_h is the number of households within post-stratum h ($h = 1, 2, ..., H$). Here, $\sum_{h=1}^{H} n_h = 32{,}425$.

Ratio. In addition to purchase totals, there are three characteristics to be estimated: share, buying rate, and penetration rate. Each of these three characteristics can be expressed as the ratio of two different totals, $R = X^{-1} Y$:

- share (X = total product category purchase and Y = item-level total purchase);
- buying rate (X= total purchasing households for each product and Y = item-level total purchase); and
- penetration rates (X = total household and Y = total households purchasing a product).

Then, estimates for the three characteristics can also be expressed as the ratio of two total estimates:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \qquad (4)$$

where both estimators $\hat{X}$ and $\hat{Y}$ can be obtained from (1).

Using Taylor series linearization, we can then obtain variance estimates for these ratio estimators:

$$v(\hat{R}) = \hat{X}^{-2} v(\hat{Y}) - 2\hat{Y}\hat{X}^{-3} cov(\hat{Y},\hat{X}) + \hat{Y}^2 \hat{X}^{-4} v(\hat{X}) \qquad (5)$$

where $v(\hat{Y})$ and $v(\hat{X})$ can be obtained from (3), and we can also obtain the covariance estimator from the same formula except using the cross product instead of the square.

We choose to use SUDAAN® to calculate variance estimates because it incorporates survey weights with stratification with relative ease (Shah, Barnwell, and Bieler 1996).

4. Design options

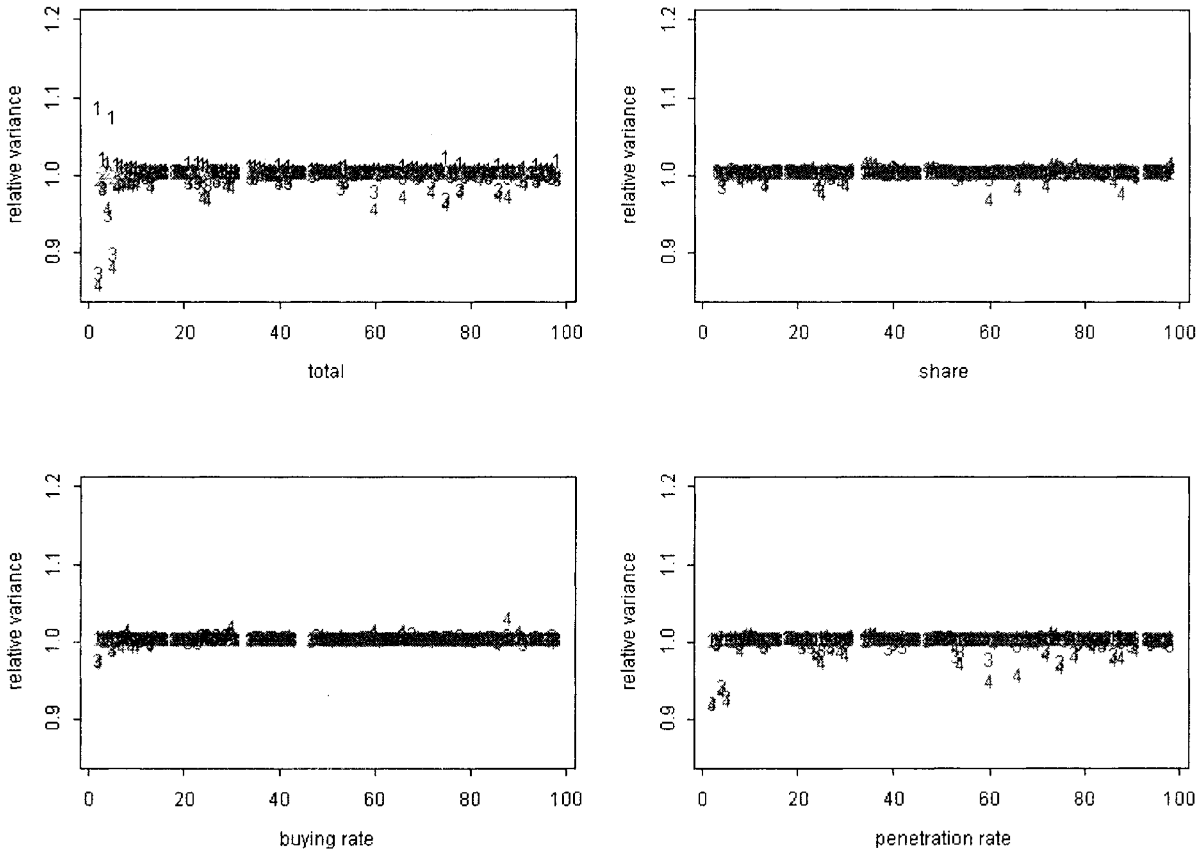
It is quite tedious, and perhaps unrealistic, to produce variance estimators that explicitly account for the raking adjustments. Consequently, we have considered simple design options that ignore the raking approximation, and assume simple poststratification adjustments within selected poststrata. Specifically, we considered five simple design options before selecting one to calculate variance estimates:

- Option 0: Simple stratified sample design using 20 geographic areas (20 strata)
- Option 1: No stratification
- Option 2: Poststratification with four Nielsen county sizes within 13 market areas and 4 remaining U.S. areas; the remaining 3 markets have only one county (71 poststrata)
- Option 3: Poststratification with four household sizes by four female head age classes within each of the 20 geographic areas (320 total poststrata)
- Option 4: Poststratification with four household sizes by four income classes by four female head age within each of the 20 geographic areas; total poststrata (1,280 total postrata)

5. Results

We investigated variance estimates based on these five predetermined design options. Figure 1 presents the relative variances as the ratios of the variances based on options 1 through 4 to the variance based on option 0. As we can observe, we see only small or moderate variation in the variance estimates among the different design options empirically studied. Since Option 0 provides relatively stable variance estimates across all statistics for all domains, we chose to employ design option 0 to calculate the variance estimates. We should finally note that the work illustrated here is part of a larger project designed to develop generalized variance estimates, hence the need for initial variance estimates for several characteristics.

FIGURE 1: RELATIVE VARIANCES OF SHAMPOO PRODUCTS FOR NATION AS A WHOLE



Relative Variance # = Variance based on Option # /Variance based on Option 0

Option 0 = Stratification with 20 geographic areas
Option 1 = No stratification
Option 2 = Stratification with 4 county size groups within 20 geographic areas
Option 3 = Stratification with 4 household size by 4 family head age groups within 20 geographic areas
Option 4 = Stratification with 4 household size by 4 income by 4 family head age groups within 20 geographic areas

REFERENCES

Deville, J-C and C-E Särndal. "Calibration Estimators in Survey Sampling." *Journal of American Statistical Association*, 1992, Vol. 87, No. 418, pp376-382.

Kalton, G. and D.S. Maligalig. "A Comparison of Methods of Weighting Adjustments for Nonresponse." *Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference,* 1991, pp.409-428.

Oh, H.L. and F. Scheuren. "Weighting Adjustments for Unit Nonresponse," in *Incomplete Data in Sample Surveys*, Volume 2: Theory and Bibliographies (W.G. Madow, I. Olkin, and D. Rubin, eds.), New York: Academic Press, 1983, pp.143-184.

Shah, B.V., B.G. Barnwell, and G.S. Bieler. *SUDAAN User's Manual*, Release 7.0. Research Triangle Park, NC: Research Triangle Institute, 1996.