POSTSTRATIFICATION OF POOLED SURVEY DATA

Mansour Fahimi, Westat, Inc.

Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Composite Weights, Design Effect, Poststratification, Replicate Weights, Sampling Errors

0. Introduction

There are survey situations in which two independent samples are both representative of the same target population, yet each sample is selected from its respective frames using a specific sampling scheme. For instance, one sample could consist of a panel selected from the list of respondents to a previous survey, while the second could be a new cross-sectional sample. In a telephone survey, one sample could be obtained from a list, while the second could be obtained through random digit dialing (RDD). It is usually the case that for analysis survey data from the two samples are to be combined (pooled).

Alternatively, there are state-level surveys that are conducted independently of the national surveys, where both surveys use similar instruments. In these situations, one might be interested in pooling survey data from a particular state with those obtained from the corresponding subset of the national survey. Some examples include: Assessment of Educational Progress, Adult Literacy Survey, Health Interview Survey, and UK Fitness Survey. All these surveys have national as well as independent state-level components. This work reviews the conventional method for pooling data for such surveys using composite weights, and presents an alternative weighting method for situations fitting the preceding descriptions.

1. Statement of the Problem

Consider a population of N units from which two independent samples of size n_1 and n_2 have been selected. Upon completion of these surveys, data from the two samples are to be pooled and weighted such that the resultant sample of size $n = n_1 + n_2$ would aggregate to the target population of N sampling units.

2. Conventional Composition Procedure

In order to blend the two samples in such a way that unbiased estimates could be obtained from the combined sample, the conventional method usually includes the following steps:

- Adjustment of base weights in each sample (separately) to reduce the effect of differential nonresponse;
- Poststratification of nonresponse-adjusted base weights in each sample (separately) to the target population counts; and
- Composition of the two sets of poststratified weights in some optimal manner.

Following the above steps, the resulting n final weights (actually, the subset of n corresponding to the responding sampling units) could be used to obtain unbiased estimates.

2.1 Composite Weights

When there are two independent samples and the common parameter of interest is, say, the population mean $\bar{Y}$, then the general composite estimator would have the following form:

$$\bar{y} = \alpha\bar{y}_1 + (1-\alpha)\bar{y}_2 \qquad (1)$$

where $\bar{y}_1$ and $\bar{y}_2$ represent estimates of $\bar{Y}$ as obtained from the first and second samples, respectively. Here, α is the blending or composition factor whose optimal value would be obtained by minimizing the following quantity:

$$MES(\bar{y}) = V(\bar{y}) + B^2(\bar{y}) \qquad (2)$$

where

$$V(\bar{y}) = \alpha^2 V(\bar{y}_1) + (1-\alpha)^2 V(\bar{y}_2) \qquad (3)$$

and

$$B(\bar{y}) = \alpha B(\bar{y}_1) + (1-\alpha) B(\bar{y}_2) \qquad (4)$$

are the variance and bias of $\bar{y}$, which in turn depend on the corresponding quantities for $\bar{y}_1$ and $\bar{y}_2$.

2.2 Optimal Composition

Depending on whether the two sample estimates of $\bar{Y}$ are biased or not, the optimal value of composition factor α would be defined differently. When only one of the two samples can provide unbiased estimate, for example, when:

$$B(\bar{y}_1) = B \text{ and } B(\bar{y}_2) = 0 \tag{5}$$

(a common situation in small area estimations) then the optimal value of α would be given by:

$$\alpha_{opt} = \frac{V(\bar{y}_2)}{MSE(\bar{y}_1) + V(\bar{y}_2)} \tag{6}$$

On the other hand, when it is reasonable to assume that both samples can provide unbiased estimates of the parameter of interest, that is, when:

$$B(\bar{y}_1) = B(\bar{y}_2) = 0 \tag{7}$$

then

$$\alpha_{opt} = \frac{\frac{\delta(\bar{y}_2)}{n_2}}{\frac{\delta(\bar{y}_1)}{n_1} + \frac{\delta(\bar{y}_2)}{n_2}} \tag{8}$$

where $\delta(\bar{y}_1)$ and $\delta(\bar{y}_2)$ represent the design effects associated with $\bar{y}_1$ and $\bar{y}_2$.

Moreover, there are situations where it is fair to assume that:

$$\frac{\delta(\bar{y}_1)}{\delta(\bar{y}_2)} \cong 1 \tag{9}$$

In such situations the optimal value of α is simply a function of sample sizes, that is,

$$\alpha_{opt} \cong \frac{n_1}{n_1 + n_2} = \frac{n_1}{n} \tag{10}$$

While theoretically straightforward, the conventional composition procedure entails a number of operational hurdles. First, this method requires computation of two sets of final weights, which eventually have to be composited. Second, in case replicate sampling weights are to be computed for estimation of sampling errors, the above procedure has to be repeated as many times as there are replicate groups.

In addition to the computational burden, these steps could have confounding effects on estimates of sampling errors. In order to avoid these niceties, one can use the following alternative method instead.

3. Alternative Weighting Method

Without loss of generality assume that there is only one stratum for poststratification purposes, and let:

B_{1i}: Sampling base weights from the first sample, i = 1, ..., n_1

B_{2j}: Sampling base weights from the second sample, j = 1, ..., n_2

Based on the conventional method, once poststratified, the above base weights would have the following form:

$$BP_{1i} = B_{1i} \times \frac{N}{\sum_{i=1}^{n_1} B_{1i}}$$

$$BP_{2j} = B_{2j} \times \frac{N}{\sum_{j=1}^{n_2} B_{2j}} \tag{11}$$

Finally, these weights have to be composited. Assuming that the condition in (9) holds, conventional composite weights would be computed as follows:

$$BPC_{1i} = BP_{1i} \times \frac{n_1}{n} = B_{1i} \times \frac{N}{\sum_{i=1}^{n_1} B_{1i}} \times \frac{n_1}{n}$$

$$BPC_{2j} = BP_{2j} \times \frac{n_2}{n} = B_{2j} \times \frac{N}{\sum_{j=1}^{n_2} B_{2j}} \times \frac{n_2}{n} \tag{12}$$

Now, suppose that the two samples are pooled prior to the poststratification, and then poststratified simultaneously to the target population total. In this case the alternative final weights would be given by:

$$BP_{1i}^{*} = B_{1i} \times \frac{N}{\sum_{i=1}^{n_1} B_{1i} + \sum_{j=1}^{n_2} B_{2j}}$$

$$BP_{2j}^{*} = B_{2j} \times \frac{N}{\sum_{i=1}^{n_1} B_{1i} + \sum_{j=1}^{n_2} B_{2j}} \qquad (13)$$

However, it is often the case that the two sets of base weights add up to vastly different numbers. Consequently, the above simultaneous poststratification can produce final weights with artificially inflated variances. Hence, these weights have to be calibrated such that they are combinable (normalized to a unified scale). The procedure described next adjusts the base weights so that they can be combined and poststratified simultaneously.

4. Calibration of Base Weights for Simultaneous Poststratification

It would be desirable if the alternative weighting procedure produces final weights that are identical to those obtainable via the conventional weighting procedure, that is,

$$\begin{cases} BPC_{1i} \equiv BP_{1i}^{*}, \forall i \\ BPC_{2j} \equiv BP_{2j}^{*}, \forall j \end{cases} \qquad (14)$$

The above conditions would hold if the following is satisfied:

$$\begin{cases} \dfrac{n_1 NB_{1i}}{n\sum_{i=1}^{n_1} B_{1i}} = \dfrac{NB_{1i}}{\sum_{i=1}^{n_1} B_{1i} + \sum_{j=1}^{n_2} B_{2j}}, \forall i \\ \dfrac{n_2 NB_{2j}}{n\sum_{j=1}^{n_2} B_{2j}} = \dfrac{NB_{2j}}{\sum_{i=1}^{n_1} B_{1i} + \sum_{j=1}^{n_2} B_{2j}}, \forall j \end{cases} \qquad (15)$$

That is:

$$\begin{cases} BPC_{1i} \equiv BP_{1i}^{*}, \forall i \\ BPC_{2j} \equiv BP_{2j}^{*}, \forall j \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n_1} B_{1i} = n_1 \\ \sum_{j=1}^{n_2} B_{2j} = n_2 \end{cases} \qquad (16)$$

The above derivation shows that instead of separately poststratifying base weights from the two samples and then producing composite weights, one can use the proposed calibration procedure to adjust base weights from the two samples such that the two could be combined and poststratified simultaneously. Specifically, base weights from each of the two samples first have to be normalized with respect to their corresponding sample sizes. Having done this, the calibrated base weights from the two samples can be pooled and poststratified concurrently.

It should be noted that the proposed calibration procedure easily carries over to more realistic situations where there are more than one poststrata. In such situations, the underlying assumption in (9) is not as restrictive as it seems. Nonetheless, even if the condition in (9) is far-fetched, one can apply the above procedure under the less restrictive condition in (8). Here, the preceding calibration would require:

$$\begin{cases} BPC_{1i} \equiv BP_{1i}^{*}, \forall i \\ BPC_{2j} \equiv BP_{2j}^{*}, \forall j \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n_1} B_{1i} = \dfrac{n_1}{\delta(\bar{y}_1)} \\ \sum_{j=1}^{n_2} B_{2j} = \dfrac{n_2}{\delta(\bar{y}_2)} \end{cases} \qquad (17)$$

That is, when the design effects of $\bar{y}_1$ and $\bar{y}_2$ do not ratio to unity, then the corresponding base weights have to be normalized with respect to the effective sample sizes. Virtually in all situations, reasonable estimates of design effects are obtainable either through approximation or use of stable estimates from previous surveys.

5. Concluding Remarks

There are clear operational gains to be expected when using the alternative weighting method. As mentioned earlier, this gain becomes more attractive when replicate weights have to be produced, where the conventional process of composition has to be repeated for each replicate separately. On the other hand, pooling survey data prior to poststratification increases the sample size, thereby allowing for finer adjustment cells. Moreover, bypassing the various computational steps needed for composition of full sample and replicate weights, along with a more elaborate poststratification, is bound to produce more stable estimates of sampling errors.