Marketing
Systems
Group

THE MATHEMATICS OF SAMPLE MIXOLOGY EFFICIENT METHODOLOGY FOR POOLING DATA FROM MULTIPLE SURVEYS

Mansour Fahimi, Ph.D. Marketing Systems Group

Abstract

In light of growing challenges for traditional methods of sample surveys, such as frame undercoverage and mounting rates of nonresponse, practitioners are increasingly compelled to consider innovative alternatives for sampling and weighting applications. What further promote such innovations are the escalating costs of conventional methods of data collection, on the one hand, and availability of data through less expensive options, on the other. As such, an emerging alternative for survey sampling is one that relies on combining multiple samples to increase the size of inferential base in a defensible and cost-effective manner (Fahimi 2015). This work provides an overview of the traditional method used for pooling data from multiple independent surveys via composite estimation and offers an efficient alternative that is more stable and computationally less cumbersome. The proposed methodology is of particular utility for the many instances where probability and nonprobability samples are combined to reduce cost, or to deal with surveys of rare subgroups.

Key Words: Composite Estimation, Design Effect, Probability and Nonprobability Samples.

Introduction

Increasingly, surveys rely on more than one sample source to improve coverage or secure the needed sample size in a cost-effective manner. Oftentimes, two or more independent samples are selected from separate sampling frames with varying representations of the population of interest. For instance, there could be two probability sample surveys with one relying on dual-frame RDD and another on address-based sampling methodologies. Alternatively, one sample could be selected from an incomplete online panel while the second could be a representative sample from a complete frame. Finally, there are many instances where various nonprobability samples are combined to create a larger base for reporting purposes. To enhance the inferential possibilities in such instances, survey data from different samples are combined prior to analysis.

Data pooling is also relevant to regional surveys that are conducted independently of national surveys, but in which both surveys collect identical data. In these situations, one might be interested in combining data from a regional survey with those obtained from the corresponding subset of the national survey. Some examples include National Assessment of Educational Progress (NAEP), National Assessment of Adult Literacy (NAAL), and the Behavioral Risk Factor Surveillance System (BRFSS). These surveys have national as well as independent subnational components that can be combined in an optimal fashion to produce estimates with improved precision at the overlapping subnational levels.

Traditionally, the conventional method of Composite Estimation has been used to mix results from different surveys to improve the robustness of the resulting estimates. That is, instead of pooling disaggregated data from different surveys and producing estimates from the combined data, individual point estimates from different surveys are produced and then blended together – one estimate at a time (Cochran 1977). In the next section the mathematical foundation for this arduous approach is reviewed, after which a more efficient alternative is introduced that can produce more stable estimates while reducing computational burden.

Mathematical Foundation

For illustration purposes, the following development will focus on two surveys. As seen later, however, the proposed methodology can easily extend to multiple surveys. As such, consider a population of N units from which two independent samples of size n_1 and n_2 have been selected. Under the conventional composition methodology, estimates from the two samples are combined to produce composite estimates that might be more robust. When the parameter of interest is, say population mean $\bar{Y}$, the general composite estimator will have the following form:

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

In the above, $\bar{y}_1$ and $\bar{y}_2$ represent estimates of $\bar{Y}$ as obtained from the first and second samples, respectively (Hansen, Hurwitz, and Madow 1953). Subsequently, an optimal value for the blending or composition factor α can be obtained by minimizing the mean square error of $\bar{y}$:

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

In the above formulation, the variance and bias of $\bar{y}$, which in turn depend on the corresponding estimates obtained from the two samples, $\bar{y}_1$ and $\bar{y}_2$, will be:

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

and

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

Depending on whether the two sample estimates ($\bar{y}_1$ and $\bar{y}_2$) are biased or not, the optimal value of the composition factor α can be defined differently. Under the general scenario when neither of the two estimates can be considered unbiased, this optimal value can be obtained and decomposed into its component parts by (Levy and Lemeshow 1991):

$$\alpha_{optimal} = \frac{MSE(\bar{y}_2)}{MSE(\bar{y}_1)+MSE(\bar{y}_2)} = \frac{V(\bar{y}_2)+B^2(\bar{y}_2)}{[V(\bar{y}_1)+B^2(\bar{y}_1)]+[V(\bar{y}_2)+B2(\bar{y}_2)]}$$

In the simplest form when pooling data from two probability samples, it can be assumed that $B(y_1) = B(\bar{y}_2) = 0$. As such, the above reduces to:

$$\alpha_{optimal} = \frac{V(\bar{y}_2)}{V(\bar{y}_1)+V(\bar{y}_2)}$$

Furthermore, when survey estimates from the two samples are expected to have similar variabilities, the above becomes a function of the sample sizes n_1 and n_2 and design effects associated with the two estimates: $\delta(\bar{y}_1)$ and $\delta(\bar{y}_2)$. Hence, the optimal value of the composition factor can be obtained by (Kish 1965):

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

Lastly, there are situations where it is justifiable to assume that:

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

In such situations the optimal value of α reduces to a simple function of sample sizes:

$$\alpha_{optimal} \cong \frac{n_1}{n_1+n_2} = \frac{n_1}{n} \Longrightarrow \bar{y} = \frac{n_1\bar{y}_1+n_2\bar{y}_2}{n}$$

Even when every one of the above simplifying assumptions can be justified, the conventional composition procedure entails several inferential and operational inefficiencies. As mentioned earlier, this burdensome approach requires that composite estimates be produced one estimate at a time. More importantly, this piecemeal process produces estimates that are based on individual samples of size n_1 and n_2, and not the larger sample of size $n_1 + n_2$. This means relying on two sets of weights created using different methodologies with adjustment granularities that will be coarser than what would be possible with a larger combined sample. Also, in case replicate weights are to be computed for estimation of sampling errors, the above procedure must be repeated as many times as there are replicate groups (Wolter 1985).

The proposed methodology detailed next eliminates the above inefficiencies and the incommoding computational complexities by furnishing the needed mathematical machinery that would allow the two samples be combined for computing a single set of Composite Weights. Specifically, instead of producing composite estimates one at a time by computing individual composition factors, under this methodology a single set of weights will be used to generate estimates from the combined sample with all the inferential dividends the larger combined sample can offer.

Composite Weights

For ease of illustration and without loss of generality, we can assume there is only one weighting cell for poststratification purposes, and let:

- B_{1i} : Sampling base weights from sample one, i = 1, …., n_1
- B_{2j} : Sampling base weights from sample two, j = 1, …., n_2

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

$$\begin{cases} BP_{1i} = B_{1i} \times \dfrac{N}{\Sigma^{n1}_{i=1}\ B_{1i}}, i = 1,...,n_1 \\ BP_{2\mathrm{j}} = B_{2\mathrm{j}} \times \dfrac{N}{\Sigma^{n2}_{\mathrm{j}=1}\ B_{2\mathrm{j}}}, j = 1,...,n_2 \end{cases}$$

Rather than producing separate point estimates using the above two sets of poststratified weights and then combining them, if the condition in (9) holds, it would be possible to produce a single set of composite weights to allow creation of point estimates from the combined data. This can be achieved by creating composite poststratified weights for the combined data as follows, however, noting the inefficiencies that still result from poststratification of smaller samples.

$$\begin{cases} BP\mathrm{C}_{1i} = B\mathrm{P}_{1i} \times \dfrac{n_1}{n} = B_{1i} \times \dfrac{N}{\Sigma^{n1}_{i=1}\ B_{1i}} \times \dfrac{n_1}{n} \\ BP\mathrm{C}_{2\mathrm{j}} = B\mathrm{P}_{2\mathrm{j}} \times \dfrac{n_2}{n} = B_{2j} \times \dfrac{N}{\Sigma^{n2}_{\mathrm{j}=1}\ B_{2\mathrm{j}}} \times \dfrac{n_2}{n} \end{cases}$$

Now, consider an alternative when the two samples are first combined and then poststratified jointly. This would be a preferred option because one can then apply more granular weighting adjustments courtesy of a larger sample that can accommodate more comprehensive poststratification possibilities. In this case the final weights will be given by:

$$\begin{cases} BP*_{1i} = B_{1i} \times \dfrac{N}{\Sigma^{n1}_{i=1}\ B_{1i} + \Sigma^{n2}_{\mathrm{j}=1}\ B_{2\mathrm{j}}} \\ BP*_{2\mathrm{j}} = B_{2\mathrm{j}} \times \dfrac{N}{\Sigma^{n1}_{i=1}\ B_{1i} + \Sigma^{n2}_{\mathrm{j}=1}\ B_{2\mathrm{j}}} \end{cases}$$

Since the two surveys are weighted to add up to the same target total N, however, the above combined poststratification ignores the fact that the corresponding two samples could have vastly different sizes. Consequently, the resulting final weights do not reflect the higher precision associated with the one survey that has a larger sample size. The procedure described next introduces a simple technique that could be used to calibrate the base weights from the two samples prior to combining them for a joint poststratification.

Calibration of Base Weights for Combined Poststratification

It would be desirable if the alternative weighting procedure could produce final weights that are identical to the composite weights. That is:

$$\begin{cases} BPC_{1i} = BP^{*}_{1i}, \forall i \\ BPC_{2j} = BP^{*}_{2j}, \forall j \end{cases}$$

The above conditions would hold if the following is satisfied:

$$\begin{cases} \dfrac{n_1 \times N \times B_{1i}}{n\Sigma^{n1}_{i=1} B_{1i}} = \dfrac{N \times B_{1i}}{\Sigma^{n1}_{i=1} B_{1i} + \Sigma^{n2}_{j=1} B_{2j}}, \forall i \\ \dfrac{n_2 \times N \times B_{12j}}{n\Sigma^{n2}_{j=1} B_{2j}} = \dfrac{N \times B_{12j}}{\Sigma^{n1}_{i=1} B_{1i} + \Sigma^{n2}_{j=1} B_{2j}}, \forall j \end{cases}$$

Simplifying the above algebra results in:

$$\begin{cases} BPC_{1i} = BP^{*}_{1i}, \forall i \\ BPC_{2j} = BP^{*}_{2j}, \forall j \end{cases} \Rightarrow \begin{cases} \Sigma^{n1}_{i=1} B_{1i} = n_1 \\ \Sigma^{n2}_{j=1} B_{2j} = n_2 \end{cases}$$

This means the alternative method produces the same composite final weights, provided that the two sets of base weights are calibrated prior to poststratification. Specifically, base weights from each of the two samples first must be scaled to their corresponding sample sizes. Having done this, instead of separately poststratifying base weights from the two samples and then producing composite weights, one can use the proposed calibrated base weights from the two samples such that the two can be combined and poststratified concurrently.

It should be noted that the proposed calibration easily carries over to more realistic situations with more than one poststratum, where the underlying assumption in (9) is easier to satisfy. Also, one can apply the above procedure under the less restrictive condition in (7) when the design effects of $\bar{y}1$ and $\bar{y}2$ do not ratio to unity. In this more realistic situation, the corresponding base weights must be normalized to their respective effective sample sizes as shown below:

$$\begin{cases} BPC_{1i}=BP^{*}_{1i},\forall i \\ BPC_{2j}=BP^{*}_{2j},\forall j \end{cases} \Rightarrow \begin{cases} \Sigma^{n1}_{i=1}\ B_{1i} = \dfrac{n_1}{\delta(\bar{y}_1)} \\ \Sigma^{n2}_{\mathrm{j=1}}\ B_{\mathrm{2j}} = \dfrac{n_2}{\delta(\bar{y}_2)} \end{cases}$$

Estimates of design effects are often readily available, or they can be quickly approximated as a function of poststratified weights by the following formula:

$$\delta(\bar{y})=1+\frac{\Sigma_{\mathrm{i}}\ \dfrac{(W_i-\bar{\bar{\bar{W}}})^2}{n-1}}{\bar{W}2}$$

Given that the combined sample will be of a larger size, it is now possible to use the calibrated base weights as input for a final stage of poststratification using an expanded set of benchmarks. Alternatively, the same set of benchmarks could be used but with higher levels of granularity to improve the representation of the combined sample with respect to finer categories of the weighting variables. Moreover, for an expediated situation one can forego the final poststratification and simply use the calibrated final weights given by:

$$W_k^{*}=W_k\times\begin{cases} \dfrac{\dfrac{n1}{\delta 1}}{\dfrac{n1}{\delta 1}+\dfrac{n2}{\delta 2}}, k=1,\ldots..n1 \\ \dfrac{\dfrac{n2}{\delta 1}}{\dfrac{n1}{\delta 1}+\dfrac{n2}{\delta 2}}, k=n1+1,\ldots..n1+n2 \end{cases}$$

Extensions and Special Cases

An extension of the above is for situations where the final weights for more than two surveys are to be blended. In such situations, the existing weights for each of the S surveys will be adjusted by the following optimal composition factors to produce the final blended weights for the combined sample:

$$\frac{\dfrac{n_k}{\delta_k}}{\Sigma^{\mathrm{s}}_{\mathrm{k=1}}\ \dfrac{n_k}{\delta_k}}\ ,\ k=1,\ldots..,S$$

Lastly, there are many instances when probability and nonprobability samples are combined. For instance, a growing number of surveys supplement their main probability samples with a less expensive samples secured from online panels from which the resulting estimates cannot claim to be unbiased. Under such scenarios when only one of the two samples can provide unbiased estimates, say $B(\bar{y}_1)=0$ but $B(\bar{y}_2)\neq 0$, the proposed calibration can be carried out using the optimal value of α obtained by:

$$^{\alpha}optimal = \frac{V(\bar{y}_2) + B^2(\bar{y}_2)}{V(\bar{y}_1) + [V(\bar{y}_2) + B^2(\bar{y}_2)]}$$

However, oftentimes nonprobability samples are used to supplement a probability sample of a modest size from which target population benchmarks are developed. Since selection probabilities are not available for nonprobability samples, typically a pseudo design weight of one is assumed for prior to poststratification. The immediate implication of this assumption is that nonprobability sample components can carry artificially smaller design effects, which means nonprobability sample components will have inflated contributions when combined weights are computed using the above formulation.

A simple solution to the above is to decompose the total design effect for the probability sample component into that due to design weights and the residual due to poststratification. The residual design effect, which then would be used in the above for the probability sample component, will be given by:

$$\delta_{Residual} = \delta_{Overall} - \delta_{Design}$$

Concluding Remarks

Conducting credible survey research in the 21st century is an endeavor subject to evolving challenges that require thinking outside of the traditional survey sampling toolbox. The statistical machinery developed by Neyman (1934) has made it possible to make measurable inferences about target populations when samples of modest size are selected from complete sampling frames; sampling units carry known selection probabilities; and surveys achieve near-perfect rates of response. For various reasons, but most notably the growing rates of nonresponse and survey costs, many of the surveys conducted these days struggle to fulfill what the traditional survey sampling paradigm requires. While such violations are fairly common among commercial surveys where theoretical underpinnings are trumped by cost and time constraints, arguably, even large-scale government surveys are not fully exempt from such concerns (Fahimi 2014).

A strategy that is often used to deal with the rising costs of surveys is to combine two or more independent samples that are selected from separate sampling frames with varying representations of the target population. In particular, such alternatives can pay considerable dividends when survey data secured from certain sample components are significantly less expensive. In some instances, multiple sample components are explicitly called for by the design of a study, while in other situations existing data from different surveys are pooled to address the size and analytical needs of a given survey. Either way, the various sample components need to be combined to produce a single analysis database. In comparison to the traditional method of composite estimation whereby separate estimates are mixed from different surveys, the proposed data pooling methodology offers at least four distinct advantages:

1. The proposed methodology is significantly less cumbersome because it enables researchers work with a single data file and not multiple files that carry separate weights.
2. Having a combined database that is larger than any of the individual survey data accommodates more granular weighting adjustments than what might be possible with individuals surveys. This becomes particularly attractive when one of the surveys is based on a small sample size.
3. A direct byproduct of the above is that the resulting survey estimates from the pooled data will be subject to smaller sampling errors due to the larger size of the combined samples.
4. Lastly, there is something to be said about applying a singular weighting methodology when working with individual survey data. Separate weighting procedures can add extraneous variations due to applications of different benchmarks, different poststratification/raking algorithms, different weight trimming rules, etc.

References

Behavioral Risk Factor Surveillance System (BRFSS), http://www.cdc.gov/brfss/index.html.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition. New York: Wiley.

Fahimi, M. (2014). “Practical Guidelines for Dual-Frame RDD – Now that the Dust is Settling” *Survey Practice*. Vol. 7, no 2, 2014, May Issue.

Fahimi, M., F. Barlas, and R. Thomas (2015). Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. *Survey Practice*, Vol. 8, no 5, 2015, December Issue.

Hansen, M., Hurwitz, W., & Madow, W. (1953). Sample Survey Methods and Theory. NY: Wiley.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Levey, P.S. and Lemeshow, S. (1991). Sampling of Populations, Methods and Applications. NY: Wiley.

National Assessment of Adult Literacy (NAAL), https://nces.ed.gov/naal/

National Assessment of Educational Progress (NAEP), http://nces.ed.gov/nationsreportcard/.

Neyman, J., 1934. “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection.” *Journal of the Royal Statistical Society 97*:558–625.

U.K. Sports Council and Health Education Authority (1992). Allied Dunbar National Fitness Survey, Sports Council, London, U.K.

Wolter, K. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.