SAS® GLOBAL FORUM 2020 #SASGF

Paper 4894-2020

Bringing SURVEYMEANS Up to Speed: Improving This Procedure for Estimating Domain Level Ratios

Brian Simonson, Cami Sorenson, Tyler Hamashima, Michael LeFew, and Soumita Lahiri, The Lewin Group

ABSTRACT

Domain-level estimation plays an important role in the field of survey sampling statistics. The implementation of domain estimation for means, totals, and ratios that SAS® has developed is available in the SURVEYMEANS procedure. Through The Lewin Group's work on statistical audits of Medicare and Medicaid claims data, two weaknesses of PROC SURVEYMEANS have been identified. First, there is no calibration weight adjustment using auxiliary variables beyond rudimentary post-stratification. Second, the algorithm for estimating domains iterates through one domain level at a time, which can be time and resource intensive. To overcome these limitations, The Lewin Group developed the Hybrid Ratio Estimator (HRE) for use in statistical audits. The HRE is a hybrid of two well-known survey estimators: the Separate Ratio Estimator (SRE) and the Combined Ratio Estimator (CRE). The HRE allows for domain-level ratio estimation with calibration weight adjustment (Deville and Särndal, 1992) and uses domain-level statistics formulas described in the groundbreaking book *Sampling Techniques* (Cochran, 1977). Leveraging these techniques, the HRE produces results for all domain levels in one iteration, which leads to vast time and resource improvement compared to PROC SURVEYMEANS. Using real-world Medicare audit data, the HRE was over seven times faster when estimating 80 domains, with time efficiency increasing drastically as the number of domain levels increased.

INTRODUCTION

Domain level estimation plays an important role in the field of survey sampling statistics. The formulas for the domain level variance of PROC SURVEYMEANS leverages an inefficient, iterative approach to produce standard errors for every domain. In this paper, we propose an alternative method of calculating standard errors for domains. For purposes of illustration, we will present a calibrated ratio estimator used in Medicare and Medicaid statistical audits. This estimator leverages a technique that is described in Cochran's (1977) *Sampling Techniques*. This variance estimation approach leads to multi-fold processing time improvement when compared to PROC SURVEYMEANS.

DOMAIN LEVEL ESTIMATION

Domain level estimation is a valuable tool to produce ratio estimates and standard errors for a specified subset of a sample of data. For example, consider a simple random sample of medical claims, where *x* is the payment amount for the claim. The projected total payment amount is given by:

$$\hat{t}_x = \frac{N}{n}\sum_{i=1}^{n} x_i$$

Now consider the projected payments for the subset of power wheelchair claims. We define the domain *x′* as follows:

$$x' = \begin{cases} x \; if \; i \in P \\ 0 \; otherwise \end{cases}$$

Where P is the space containing all power wheelchair claims. Then the projected total payment for x' is simply:

$$\hat{t}_x' = \frac{N}{n}\sum\nolimits_{i=1}^{n} x_i'$$

The variance of this projection is given by:

$$Var(\hat{t}_x') = \frac{N^2}{n} s_{x'}^2 = \frac{N^2}{n}\frac{\sum_{i=1}^{n}(x_i' - \bar{x}')^2}{n-1}$$

The formula above is the correct variance for the domain level estimate. In practice, by effectively turning all of the data not in the domain of interest into zeros, we achieve the desired estimate and correct standard error. The zeros serve as a statistical bookmark to track the random variable of the number of observations within a sample that fall within a given domain. Cochran (1977) wrote that this simple trick was one of the more powerful intuitive concepts he could instill on his students. While the simplicity of this approach is very educational, it is inefficient when used in practice. For every domain, a "zeroed out" variable must be constructed before the point estimate and standard error are calculated. However, leveraging basic algebraic identities, we transform the variance formula above into:

$$Var(\hat{t}_x') = \frac{N^2}{n}\frac{1}{n-1}\left((n^p - 1)s_{x_p}^2 + n^p\left(1 - \frac{n^p}{n}\right)\bar{x}_p^2\right)$$

Where p subscript denotes all observations within the domain, and n^p is the number of claims within the power wheelchair sample domain, P. This formula uses basic statistics of the domain as well as the overall sample to produce an estimate and standard error. These domain level statistics can be requested in one step for multiple domain levels.

PROC SURVEYMEANS AND THE HYBRID RATIO ESTIMATOR

This paper focuses on the comparison between SAS's PROC SURVEYMEANS and the Hybrid Ratio Estimator (HRE) from both a methodological and variance estimation perspective. SAS's implementation of domain estimation for means, totals, and ratios is available in PROC SURVEYMEANS. As described in the SAS documentation, to perform domain level estimation, PROC SURVEYMEANS calculates the ratio estimate and standard error using the "zeroing out" method for each domain.

The Hybrid Ratio Estimator (HRE) is a hybrid between two well-known survey estimators: the Separate Ratio Estimator (SRE) and the Combined Ratio Estimator (CRE). The CRE leverages sampling weights to project the numerator and denominator of a ratio. A simple expression for this is:

$$\hat{R}_{CRE} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i y_i}$$

where x is the numerator, y the denominator of interest, and w the associated weight for observation i. The SRE takes advantage of known universe denominator totals available in mutually exclusive sub-populations or partitions[1]. The SRE is a weighted combination of CREs, using these known totals for weighting. Continuing our simple example, if you know the values of y at the universe level by some arbitrary level M, then the SRE is:

$$\hat{R}_{SRE} = \sum_{j=1}^{M} t_y^j \hat{R}_{CRE}^j$$

[1] Partitions are unrelated to the specifications of a domain.

where t_y^j is the partition level total for quantity *y* in partition *j*, and $\hat{R}_{CRE}^{j}$ is a domain level CRE estimate for partition *j*. Benchmarking to known universe totals both serves as a powerful variance reduction tool as well as a pragmatic function where projected denominators sum to known totals.

Inspecting the formulas for the CRE and the SRE, the primary benefit of each estimator is clear. The CRE allows for domain level estimation; while, the SRE benchmarks to known universe totals. However, the SRE by definition does not allow for domain level estimation. It is the disconnect between these two estimators that led to the formulation of the HRE.

The HRE is a generalized ratio estimator that combines the form of the CRE and the SRE. The HRE allows for domain level ratio estimation with a calibrated weight adjustment (Deville and Särndal, 1992) and utilizes domain level statistical formulas described in the groundbreaking book *Sampling Techniques* (Cochran, 1977). The form of the HRE is expressed below:

$$\hat{R}_{HRE}^{d} = \frac{\hat{t}_x^{*d}}{\hat{t}_y^{*d}} = \frac{\sum_i \hat{t}_x^{*di}}{\sum_i \hat{t}_y^{*di}} = \frac{\sum_i \frac{\hat{t}_x^{di}}{\hat{t}_y^{i}} t_y^{*i}}{\sum_i \frac{\hat{t}_y^{di}}{\hat{t}_y^{i}} t_y^{*i}}$$

where *d* is the domain of interest, *i* is the partition, and *x* and *y* are the numerator and denominator of the ratio, respectively. The form above is irrespective of the specific design of the study. Further, the generalized form of the variance is given by:

$$Var\left(\hat{R}_{HRE}^{d}\right) = \frac{1}{\left(t_y^{*d}\right)^2} \sum_i \left(\frac{t_y^{*i}}{t_y^{i}}\right)^2 Var\left(\hat{\theta}^{di} - \xi^{i} \hat{t}_y^{i}\right)$$

where

$$\xi^{i} = \frac{\hat{t}_x^{di} - R^{d} \hat{t}_y^{di}}{\hat{t}_y^{i}}$$

and

$$\hat{\theta}^{di} = \hat{t}_x^{di} - R^{d} \hat{t}_y^{di}$$

Under a one-stage stratified simple random sample, the total estimator for any variable, *z*, is given by:

$$t_z^{di} = \sum_{k=1}^{a} \frac{N_k}{n_k} \sum_{j=1}^{n_k^{di}} z_{kj}$$

$$\hat{t}_z^{di} = \sum_{h=1}^{b} W_h \sum_{g=1}^{m_h} \sum_{k=1}^{a} \frac{N_{hgk}}{n_{hgk}} \sum_{j=1}^{n_{hgk}^{di}} z_{hgkj} = \sum_{h=1}^{b} W_h \hat{t}_{x_h}^{di} = \sum_{h=1}^{b} W_h \sum_{g=1}^{m_h} \hat{t}_{x_{hg}}^{di}$$

where

k denotes the stratum (*k*=1 to *a*)

j denotes the sampled unit within stratum *k*

N_k denotes the number of units in the universe of stratum *k*

n_k denotes the number of units sampled from stratum *k*

n_k^{di} denotes the number of units sampled from stratum *k*, domain *d*, and partition *i*

Then, the key term in the variance of the HRE can be shown to be:

$$Var\left(\hat{R}_{HRE}^{d}\right) = \frac{1}{\left(t_y^{*d}\right)^2} \sum_i \left(\frac{t_y^{*i}}{t_y^i}\right)^2 \sum_k \frac{N_k^2}{n_k} s^2_{\hat{\theta}''_{kj} - \hat{\xi}^i y'_{kj}}$$

Relying on the same algebra used to express domain level statistics as discussed earlier, you can determine that:

$$s^2_{\hat{\theta}''_{kj} - \hat{\xi}^i y'_{kj}} = \frac{n_k^{di} - 1}{n_k - 1} s^2_{\hat{\theta}_{kj}^{di} - \hat{\xi}^i y_{kj}^i} - \hat{\xi}^{i2} \frac{n_k^{di} - 1}{n_k - 1} s^2_{y_{kj}^{di}} + \hat{\xi}^{i2} \frac{n_k^i - 1}{n_k - 1} s^2_{y_{kj}^i} - \frac{n_k^{di}}{n_k - 1}\left(1 - \frac{n_k^{di}}{n_k^i}\right)\left(\hat{\theta}_k^{di}\right)^2$$

$$+2\frac{n_k^{di} - 1}{n_k - 1} \hat{\xi}^i \hat{\theta}_k^{di} \left(\frac{n_k^i}{n_k} \bar{y}_k^i - \bar{y}_k^{di}\right) - \frac{n_k^i}{n_k - 1}\left(1 - \frac{n_k^i}{n_k}\right) \hat{\xi}^{i2} \left(\bar{y}_k^i\right)^2$$

This formula allows for variance estimation of domains in a non-iterative fashion. While the HRE was developed to overcome the shortfalls of the CRE and SRE, the use of these proposed domain level variance calculations leads to more operational efficiencies that SAS currently does not leverage in PROC SURVEYMEANS.

PROCEDURE COMPARISON

Before performing a series of speed tests, there are functional differences between the HRE and PROC SURVEYMEANS that should be addressed including: the calibration, estimator utilized, and variance calculation.

CALIBRATION

Calibration is a widely used technique to benchmark projections to known universe totals. The universe totals need not be aligned with the strata in the sampling design and can be specified by any dimension. Calibration estimators are available in a variety of statistical software packages but is not currently available in PROC SURVEYMEANS. The post-stratification option, whereby known universe counts can be applied post sampling, is available in PROC SURVEYMEANS, but this is not recalibration. Thus, in order to ensure speed tests are performed on similarly requested estimates, no calibration was specified for the HRE in this example.

ESTIMATORS

PROC SURVEYMEANS produces sums, means, and ratio estimates; the HRE currently produces only ratio estimates. Therefore, only the speed of ratio estimation was compared for this study.

VARIANCE CALCULATION

As discussed above, the HRE and PROC SURVEYMEANS utilize different computational methods of estimating the variance for a domain. However, the HRE and PROC SURVEYMEANS produce the same domain ratios and standard errors when used with similar specifications.

PERFORMANCE COMPARISON

We performed a series of speed tests for the HRE and PROC SURVEYMEANS. Using real-world Medicare audit data, the HRE was over 7 times faster when estimating the results of 80 domains, with time efficiency increasing drastically as the number of domain levels increased. This improved performance is largely due to the mathematical efficiency of using summary statistics of the domain to produce correct domain level standard errors. With PROC SURVEYMEANS having to loop through each level within a domain to produce an estimate and standard error, the time to complete the task climbs exponentially. Table 1 shows the time (in seconds) that it took each procedure to produce ratios by the dimension(s) listed.

Dimension(s)	Number of Levels	Hybrid Ratio Estimator Run Time (in seconds)	PROC SURVEYMEANS Run Time (in seconds)
Cluster	29	3.9	17.9
Provider Type	80	4.6	33.4
Type of Service	230	4.6	110.5
Cluster & Provider Type	906	6.5	370.7
Cluster, Provider Type & Type of Service	7,793	9.0	1,801.4

Table 1. Hybrid Ratio Estimator versus PROC SURVEYMEANS Time Comparison (in seconds)

As displayed in Table 1, the HRE's vast improvement in speed compared to PROC SURVEYMEANS shows the power of using a non-iterative approach to domain estimation, especially as the number of requested domains increases.

CONCLUSION

We developed the HRE to combine the benefit of two well-known estimators: the CRE and the SRE. During the formulation of the estimator and its standard error, a non-iterative expression of the standard error was developed that leverages the basic algebraic identities of variance. This method of variance calculation provides a multifold improvement on time requirements for estimation. While we believe the HRE stands on its own merits as a generalized ratio estimator, the standard error calculations provide clear advantages for its use as a non-iterative approach to domain level estimation.

REFERENCES

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

Deville, J.C. and Sarndal, C.E. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association*, 87: 376-382.

Simonson, B. 2007. "The Hybrid Ratio Estimator." *Proceedings of the ICES-III*, 754-757. Montreal, Quebec, Canada: American Statistical Association.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Brian Simonson
The Lewin Group
brian.simonson@lewin.com

Cami Sorenson
The Lewin Group
cami.sorenson@lewin.com