A Comparison of Variance Estimates Using Random Group and Taylor Series Methods for a Large National Survey of Businesses

Sadeq R. Chowdhury and David Kashihara
Agency for Healthcare Research and Quality[1]
5600 Fishers Lane, Rockville, MD 20857

Abstract

The random group (RG) method of variance estimation was used in the Medical Expenditure Panel Survey – Insurance Component (MEPS-IC) from the beginning of the survey in 1996 through 2013. This method was found to be less reliable for certain types of estimates so the variance estimation method was changed to the Taylor Series Expansion (TS) method starting in 2014. This paper presents a comparison of standard error estimates using the RG and TS methods for a variety of MEPS-IC estimates and will describes situations where the RG method may be less reliable than the TS method.

Key Words: MEPS, Random Group, Taylor Series, Variance

1. Introduction

The Medical Expenditure Panel Survey – Insurance Component (MEPS-IC) is an annual survey of private employers as well as state and local governments that has been conducted since 1996. The survey produces national and state-level estimates of employer-sponsored health insurance including estimates of the number of offered plans, the number of enrolled employees, and items such as health insurance premiums, copayments, and deductible amounts. The MEPS-IC is sponsored by the Agency for Healthcare Research and Quality (AHRQ) and is fielded by the U.S. Census Bureau. The annual private-sector sample is comprised of roughly 42,000 business establishments. An establishment is a single business entity or location as opposed to a firm, also known as a company, which can comprise one or more establishments. Government agencies in the MEPS-IC include all state governments including the District of Columbia, as well as a sample of local governments. A sampled government includes the parent agency and all of the dependent agencies that are associated with that parent agency. Annually there are about 3,000 state and local governments sampled in the MEPS-IC (Davis, 2015).

[1] The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services (DHHS) or the Agency for Healthcare Research and Quality (AHRQ) are intended or should be inferred.

For the private sector, the list-based sampling frame is the Census Bureau’s Business Register. The public sector sample is selected from the Census Bureau’s Governments Integrated Directory (GID). The sampling is a stratified, single-stage sample of establishments or government agencies using equal probability sampling (EPS) for the private sector and probability proportional to size (PPS) sampling for the public sector with clustering of plans within establishments and government units. MEPS-IC estimates are available in tables on the MEPS web site (https://meps.ahrq.gov/mepsweb/) and can also be generated using MEPSnet, also found on the web site. MEPSnet enables users to produce estimates at different levels of aggregation as well as data trends for survey years 1996 to 2016. Researchers may also apply to use the restricted-access microdata at designated Research Data Centers such as the ones located at the U.S. Census Bureau in Suitland, Maryland and at the University of Maryland in College Park, Maryland.

The Random Group methodology (Wolter, 1985) was historically used to produce estimates of variance for the MEPS-IC. This method was used by other surveys at the Census Bureau so incorporation into a new survey was relatively easy. However, as the number of published tables and stub variables within those tables grew over time, some technical shortcomings became evident with the methodology. Thus, starting in 2014, the variance estimation methodology was changed from Random Group (RG) to Taylor Series linearization (TS). This paper describes the technical shortcomings of RG variance estimates and compares RG estimates, in terms of relative standard errors (RSEs), with those calculated based on the TS method using 2013 MEPS-IC data for the purpose of illustrating reasons for the change in variance estimation methodology.

2. Variance Estimation Methods used by the MEPS-IC

The two variance estimation methods employed by the MEPS-IC each have their own merits. The methods are described in the paragraphs below and are then followed by a comparison of various types of estimates to illustrate the benefits of using the Taylor Series linearization method in the MEPS-IC.

2.1 Random Group (RG) Method

The Random Group estimator for the variance of an estimate, $\hat{\theta}$, is computed as

$$var(\hat{\theta})=\frac{1}{k(k-1)}\sum_{\alpha}^{k}\left(\hat{\theta}_{\alpha}-\hat{\theta}\right)^{2}$$

where k is the number of random groups, $\hat{\theta}$ is the estimate based on the entire sample, and $\hat{\theta}_{\alpha}$ is the estimate based on the establishments in random group α.

During the sequential sample selection process in the MEPS-IC, each selected establishment is assigned a number corresponding to its place in the order of selection. These selection numbers are converted to $\alpha = 10$ groups numbered 0 to 9 by assigning an establishment to the group determined by the last digit in its selection number. Thus, if the selection number were 73, the establishment would be assigned to group 3. Each group can then be thought of as a subsample similar to the full sample with each unit having a chance of selection into the subsample that was one-tenth its chance of selection into the full sample. Using subsample weights that are 10 times the nonresponse adjusted weights of the full sample, 10 subsample estimates, $\hat{\theta}_\alpha, \alpha = 0, 1, 2, \cdots 9$ are made in addition to the full sample estimate, $\hat{\theta}$. These values of $\hat{\theta}_\alpha$ and $\hat{\theta}$ are used in the above formula for computing the RG variance of the estimate, $\hat{\theta}$.

2.2 Taylor Series (TS) Method

Beginning with 2014 data, standard errors in MEPS-IC are computed using the Taylor Series linearization method which is a widely used method of variance estimation for sample survey estimates. Under the TS method, standard variance estimation formulae available for linear estimators are used for all estimators. The TS method is also used for variance estimation in the MEPS Household Component (Chowdhury, 2013). For nonlinear estimators, the linear approximation is obtained by using a first-order Taylor series expansion.

For the i-th establishment in stratum h, if x_{hi} is the value of a target variable, w_{hi} is the estimation weight (which can just be the inverse of the selection probability in the absence of any nonresponse or other adjustment), $\theta_{hi} = w_{hi}x_{hi}$ is the corresponding weighted value, and n_h is the sample size in the stratum then the variance of an estimator of the total, $\hat{\theta} = \sum_{h=1}^{H} \hat{\theta}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}x_{hi} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \theta_{hi}$, is obtained under the TS method as

$$var(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h}{(n_h-1)} \sum_{i=1}^{n_h} (\theta_{hi} - \bar{\theta}_h)^2 \text{ with } \bar{\theta}_h = \frac{\sum_{i=1}^{n_h} \theta_{hi}}{n_h}$$

The sampling rate can be high in some MEPS-IC strata. Therefore, beginning in 2016, a finite population correction factor (FPC) was incorporated into the variance calculation. Thus, with the finite population correction factor, $FPC = (1 - n_h/N_h) = (1 - f_h)$, the variance is computed as

$$var(\hat{\theta}) = \sum_{h=1}^{H} (1 - f_h) \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} (\theta_{hi} - \bar{\theta}_h)^2$$

The formulas for variance estimation using the TS method for different nonlinear estimates produced from the MEPS-IC can be found in SAS Stat User’s Guide (2012) or SUDAAN User’s Manual (1996).

The standard error (SE) and the relative standard error (RSE) of the estimate $\hat{\theta}$ is defined as

$$SE(\hat{\theta}) = \sqrt{var\ (\hat{\theta})} \text{ and } RSE(\hat{\theta}) = SE/\hat{\theta}$$

and RSEs are often expressed in percentages.

For more information about the Taylor series method of variance estimation see Cochran (1977), Lohr (2009); Särndal, Swensson, and Wretman (1992); Lee, Forthoffer, and Lorimor (1989); and Wolter (1985).

3. A Comparison of Variance Estimates

The variances computed using RG and TS methods are compared for a wide range of MEPS-IC private sector estimates using 2013 data. The estimates are compared separately for totals of continuous variables (such as premiums, contributions, enrollments, etc.) and percentages for categorical variables (such as offer rates, take-up rates, eligibility rates, etc.) both at the U.S. and State levels by firm size, industry group, age of firm, ownership, low wage, union presence and multi/single status. Variance estimates for about 38,000 estimates of totals and about 68,000 estimates of percentages are compared. The RSEs of all estimates using both RG and TS methods are produced and the percentage point differences in RSE (i.e., Diff RSE % = RG RSE % - TS RSE %) are computed for each estimate.

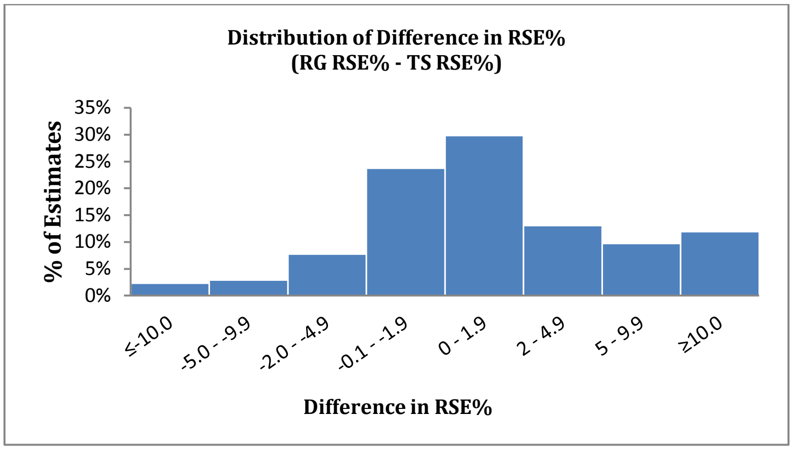
3.1 Difference in RSE Estimates of Totals

Table 1 shows the distribution of percentage point differences in estimated RSEs between RG and TS methods and Figure 1 shows the corresponding histogram of such differences for estimates of totals. Table 1 shows that the difference is less than ± 2 percentage points for about 53 percent of the estimates and less than ± 5 percentage points for about 74 percent of the estimates. However, the difference is wider than ± 5 percentage points for about 26 percent of the estimates. Table 1 also shows that for the majority of estimates (64 percent), the RSE estimate under the RG method is higher than the RSE estimate under the TS method.

Table 1: Distribution of differences in RSEs of estimates of totals computed using RG and TS methods

Difference in RSE% (RG-TS)	Number of estimates	Percent of estimates	Percent of estimates	
≤-10.0	817	2.1%		4.8%
-5.0 - -9.9	1,044	2.7%		
-2.0 - -4.9	2,894	7.6%		73.8%
-0.1 - -1.9	8,973	23.6%	53.5%	
0 – 1.9	11,296	29.7%		
2 – 4.9	4,907	12.9%		
5 – 9.9	3,645	9.6%		21.4%
≥10.0	4,482	11.8%		
Total	38,058	100%		100%

Figure 1: Histogram of difference in RSEs of estimates of totals computed using RG and TS methods



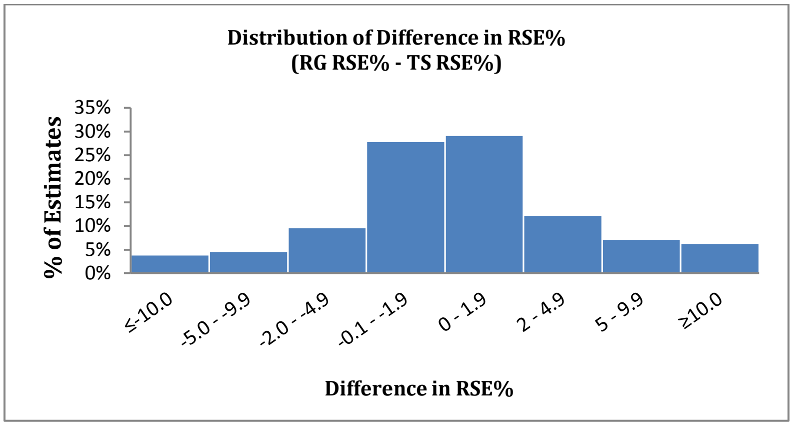
3.2 Difference in RSE Estimates of Percentages

Table 2 presents the distribution and Figure 2 presents the corresponding histogram of the differences in estimated RSEs for about 68,000 percentage estimates. For about 57 percent of the estimates the difference in RSEs is less than ± 2 percentage points and for about 78 percent estimates the difference is less than ± 5 percentage points. However, for about 22 percent of the estimates the difference is wider than ± 5 percentage points. For 55 percent of all estimates, the RSE estimate under the RG method is higher than the RSE estimate under the TS method.

Table 2: Distribution of differences in RSE of estimates of percentages computed using RG and TS methods

Difference in RSE% (RG-TS)	Number of estimates	Percent of estimates	Percent of estimates	
≤-10.0	2,553	3.8%		8.3%
-5.0 - -9.9	3,069	4.5%		
-2.0 - -4.9	6,471	9.5%		78.5%
-0.1 - -1.9	18,895	27.8%	56.9%	
0 - 1.9	19,775	29.1%		
2 - 4.9	8,256	12.1%		
5 - 9.9	4,809	7.1%		13.3%
≥10.0	4,237	6.2%		
Total	68,065	100%		100%

Figure 2: Histogram of difference in RSEs of estimates of percentages computed using RG and TS methods



3.3 Effect of Sample Size on Difference in RSE Estimates

Sample size appears to be associated with differences in RSEs between the RG and TS methods. The RSE estimate using the RG method tends to be higher than using the TS method when the sample size is smaller. If the estimates of RSEs, which are based on a sample of size 50 or less, are excluded from the comparison then the difference in RSEs appears to be very small. For the RSEs of the estimates of totals, when the difference in RSE% is ≥ 5 percentage points, 68 percent of the estimates are based on sample sizes of 100 or less (Table 3). For the RSEs of the estimates of percentages, the effect of small sample size on the difference in RSEs is even more evident. When the difference in RSE percentages is ≥ 5 percentage points, 90 percent of the estimates are based on a sample size of 100 or less (Table 4). In other words, if the estimates of RSEs of totals based on a sample of size 50 or less are excluded from the comparison, then for 69 percent of the estimates, the difference in RSEs is less than ± 2 percentage points and for 91 percent of the estimates, the difference in RSEs is less than ± 5 percentage points. For the SEs of percentage estimates, if the estimates based on a sample size of 50 or less are excluded then for 63.5 percent of the estimates, the RSE difference is less than ± 2 percentage points and for 85.2 percent of the estimates, the RSE difference is less than ± 5 percentage points.

Table 3: RSE difference by sample size (estimates of totals)

RSE Diff =|RG-TSE|	Sample Size						
	<30	30-50	50-100	100-200	200-500	500+	Total
<5	3.4%	6.7%	17.6%	15.8%	30.4%	26.1%	100.0%
≥5	24.0%	18.8%	25.3%	12.6%	13.2%	6.0%	100.0%

Table 4: RSE difference by sample size (estimates of percentages)

RSE Diff =|RG-TSE|	Sample Size						
	<30	30-50	50-100	100-200	200-500	500+	Total
<5	5.5%	6.4%	17.9%	20.7%	26.0%	23.5%	100.0%
≥5	56.6%	17.9%	15.9%	7.1%	2.3%	0.2%	100.0%

4. Discussion

Research into the validity of Random Group (RG) variance estimates in the MEPS-IC survey under certain conditions, such as empty groups, were examined by Baskin (2014) and Thompson (2014). These and other issues led to the decision to calculate survey variances using Taylor Series (TS) Linearization rather than Random Groups. RG variances are problematic when there are empty groups or unbalanced groups in small domains. Further, the finite population correction factor (FPC) is difficult to incorporate. In contrast, the FPC can easily be incorporated into TS variances and no group formation is necessary. Another benefit of utilizing an FPC term is that it can be used as a means to account for nonresponse in certainty strata.

When comparing RSEs for RG and TS variance estimates, high correlation was generally observed between the two methods for estimates of totals as well as for estimates of percentages. However, for about 20 percent of the cases the RSE difference was 5 percentage points or more. The largest differences were mostly in estimation cells of less than 100 cases. Also, for a majority of comparisons, the RG estimates were greater than the TS estimates. In summary, this analysis provides evidence that switching from RG to TS variance estimates is beneficial to the MEPS-IC survey for both theoretical and computational reasons.

5. Acknowledgement and Disclaimer

The authors would like to acknowledge the valuable contributions to this project by Matthew Thompson of the U.S. Census Bureau.

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

6. References

Baskin, R. (2014). "Statistical Properties of Assigning Zero to Empty Groups in Random Groups Variance Estimation". AHRQ Internal Paper.

Chowdhury, S.R. (2013). A Comparison of Taylor Linearization and Balanced Repeated Replication Methods for Variance Estimation in Medical Expenditure Panel Survey. Agency for Healthcare Research and Quality Working Paper No. 13004, July 2013, http://meps.ahrq.gov/mepsweb/data_files/publications/workingpapers/wp_13004.pdf

Cochran, W. G. (1977). *Sampling Techniques*, Third Edition, New York: John Wiley & Sons, Inc.

Davis, K. *Sample Design of the 2014 Medical Expenditure Panel Survey Insurance Component*. Methodology Report #30. June 2015. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr30/mr30.pdf

Fuller, W. A. (1975). "Regression Analysis for Sample Survey," *Sankhy*_a, 37 (3), Series C, 117–132.

Lee, E. S., Forthoffer, R. N., and Lorimor, R. J. (1989). *Analyzing Complex Survey Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-071, Beverly Hills and London: Sage Publications, Inc.

Lohr, S. L. (2009). ***Sampling: Design and Analysis***, Second Edition, Pacific Grove, CA: Duxbury Press.

Särndal, C.E., Swenson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag Inc.

Skinner, C. J., Holt, D. and Smith, T. M. F. eds. (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

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

Thompson, M. (2014). "Evaluation of Standard Errors Calculated Using Empty Random Groups". U.S. Census Bureau Internal Paper.

Woodruff, R. S. (1971). "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, 66, 411–414.

SAS Stat User's Guide (2012). SAS Institute Inc., 2012. SAS/STAT® User's Guide, Carey, NC.

SUDAAN User's Manual (1996). SUDAAN® User's Manual, Release 7.5. Research Triangle Park, NC: Research Triangle Institute.