

Incorporating a Finite Population Correction into the Variance Estimation of a National Business Survey¹

Sadeq R Chowdhury*, David Kashihara*, Matthew Thompson**

*Agency for Healthcare Research and Quality (AHRQ), 5600 Fishers Lane, Rockville, MD 20857

**U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746

Abstract

The Medical Expenditure Panel Survey (MEPS) - Insurance Component (IC) is a large national annual survey of private business establishments as well as state and local governments. It is a major source of information on employer-related health insurance in the United States. The sample for MEPS-IC is selected using a stratified design from two list-based frames. In many sampling strata the sampling rate is very high but the finite population correction (FPC) was not included in the variance estimation until 2016. This paper discusses how the FPC factor was incorporated into the Taylor Series variance estimation in various sampling strata with different types of sampling. It also illustrates the impact of incorporating the FPC on MEPS-IC variance estimates.

Key words: Finite Population Correction, Variance Estimation, Business Survey, MEPS

1.0 Introduction

The variances of estimates from sample surveys are usually computed assuming the population is infinite or the sample is selected with replacement or the sampling rate is negligible. However, surveys are often conducted based on samples without replacement from finite populations with a non-negligible sampling rate. In this case, the variances of certain survey estimates are overestimated under the usual assumptions and, to estimate variances correctly, a finite population correction (FPC) factor should be applied particularly when the sampling rate is high.

Some may argue that an FPC is not always required even if the sample is selected from a finite population. This depends on whether the inference is intended for the finite population in hand or for a wider population than the given finite population (Rust et al., 2017, Deming and Stephan, 1941, Graubard and Korn, 2002). If the interest is about a finite population in a particular point in time or about year to year change in a characteristic of the population, say for monitoring the impact of a policy change, then it is recommended to apply the FPC to compute variances more accurately.

Starting with the 2016 survey year, an FPC factor was incorporated in the process of variance estimation for the Medical Expenditure Panel Survey – Insurance Component (MEPS-IC). Including an FPC factor became more feasible when the variance estimation methodology was changed from Random Groups to Taylor Series linearization in 2014 because it is technically less involved to include an FPC factor in the Taylor Series method than in the Random Groups method. Including an FPC is expected to improve the variance estimation since the sampling fraction in some strata of the MEPS-IC can be fairly high. Also, in a certainty stratum when there is nonresponse, the inclusion probability becomes less than 1. In this case, a certainty stratum can be treated as a noncertainty stratum for variance estimation and an FPC adjustment can be used based on the response rate to determine if the variance should be close to zero or notably greater than zero. In this paper, we will discuss how the FPC factor was incorporated into the Taylor Series variance estimation methodology in various sampling strata with different types of sampling in the MEPS-IC. We will also assess the effect of incorporating the FPC factor on the variances of various survey estimates.

2.0 Medical Expenditure Panel Survey – Insurance Component (MEPS-IC)

The 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

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) or the U.S. Census Bureau is intended or should be inferred.

Healthcare Research and Quality 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 may 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 agency includes all of the dependent units that are associated with the parent agency. Annually there are about 3,000 state and local government agencies sampled in the MEPS-IC (Davis, 2015).

For the private sector, the sampling frame is the Business Register of the U.S. Census Bureau and, for the public sector, the sampling frame is the Census Bureau's Governments Integrated Directory. Stratified single-stage samples of private sector establishments are selected with equal probability while stratified public sector government agencies are selected with probability proportional to size (PPS) with all dependent units or sub-agencies within sampled government agencies are included in the sample with certainty. Health insurance plans are within sampled establishments or government sub agencies are selected. For private sector establishments, up to four health insurance plans are sampled within each establishment. If the establishment offers more than four plans, the three largest plans are selected and a fourth plan is sampled from the remaining plans. For government agencies, information on all health insurance plans is collected.

The Random Group methodology (Wolter, 1985) was historically used to produce variance estimates for the MEPS-IC. Since this method was used by other surveys at the Census Bureau at that time, it was easy to adapt and was used for the MEPS-IC. However, as the number of published tables and stub variables within those tables grew over time for the MEPS-IC, some technical shortcomings became evident with the methodology (Chowdhury and Kashihara, 2017). To address this, the variance estimation methodology was changed from Random Groups to Taylor Series linearization in 2014. Due to the change in the variance estimation methodology, the incorporation of an FPC factor became easier and was incorporated subsequently starting with 2016.

3.0 The Finite Population Correction

The FPC factor is defined as

$$FPC = \left(1 - \frac{n}{N}\right) = (1 - f)$$

where, n is the sample size, N is the corresponding population size and $f = n/N$ is the sampling fraction in a stratum (Cochran, 1977). The FPC factor equals the proportion of the population not included in the sample. It reduces the variance of survey estimates when the sampling fraction is not negligible.

The above expression for the FPC does not consider any nonresponse which is consistent with what is usually presented. However, since the MEPS-IC like other surveys is subject to nonresponse we will define FPC after allowing for nonresponse as

$$FPC = \left(1 - \frac{n_r}{N}\right)$$

where n_r is the responding sample size in a stratum².

4.0 Incorporating FPC in MEPS-IC

This section describes how the FPC factor is defined and specified in the computation process for estimating the variances of various types of MEPS-IC estimates.

4.1 Private Sector Estimates

4.1.1 Noncertainty Strata

For establishment-level estimation, the FPC factor is defined as $FPC = \left(1 - \frac{n_r}{N}\right)$, where n_r is the responding number of establishments in a stratum and N is the corresponding number of eligible establishments in the population for the same stratum. The population counts (N) are obtained by counting the number of establishments by strata on the sampling frame³.

² this essentially treats the respondents within each stratum as a random sample

³ For both Private and Government estimation, since some establishments/agencies on the frame often are found to be not eligible during the survey, it was investigated whether to use the sum of weights or frame count in a stratum as the population total in

For plan-level estimation, since an establishment serves as a primary sampling unit (PSU), the FPC is calculated based on the number of establishments, i.e., both the numerator and the denominator in the FPC are respective counts of establishments (not plans).

In some strata, the establishments are selected with PPS without replacement. For PPS sampling, the variance estimation is usually done under the assumption of with-replacement sampling where FPC is not relevant. However, for the sake of consistency, an FPC factor (as defined above) is included in the variance estimation also in strata where establishments are selected with PPS. Since the PPS sampling in the MEPS-IC is usually done in larger strata with a smaller sampling rate, the FPC adjustment is not expected to make much difference in the variance estimates. Nevertheless, even if the difference is small it should improve the variance estimate, because estimating variance under the assumption of with replacement, when the sampling was actually done without replacement, and not including an FPC overestimates the variance and the overestimation is inversely proportional to the *FPC* factor (Cochran, 1977, Sarndal et al., 1992).

4.1.2 Certainty Strata

Prior to the implementation of the FPC, the variance in a certainty stratum was forced to be zero. This would be true if all cases in a certainty stratum respond but often there is nonresponse among certainty cases that introduces an uncertainty in the estimate for sampling introduced among certainties due to nonresponse. As mentioned earlier, one of the objectives of incorporating the FPC is to use the FPC factor to determine if there will be a variance in a certainty stratum. In order to allow for the possibility of non-zero contribution to variance from certainties, under this new approach, certainties are grouped into variance strata and treated as noncertainty strata while FPCs within strata are used to determine whether the variance should be zero or not. When all certainty establishments within a certainty stratum respond, the FPC will be zero and hence the variance will be zero. If one or more certainty establishments do not respond then the FPC becomes nonzero and there will be a nonzero variance for a certainty stratum. The variance strata for private sector certainties are defined using state (i.e., a total of 51 variance strata for certainties comprising 50 states plus Washington DC).

The FPC for a certainty stratum is then defined as $FPC = \left(1 - \frac{n_r}{N}\right)$, where n_r is the number of responding certainties and N is the total number of certainties in the variance stratum⁴. Similar to the noncertainty strata, the same FPC is used for both establishment and plan level estimation because an establishment is the PSU for the sampling of plans.

4.2 Government Sector Estimates

4.2.1 Noncertainty Strata

For the government sector noncertainty strata, the FPC is incorporated in the same way as for the private sector noncertainty strata as discussed above. The population counts (N) are created by counting the number of parent agencies on the frame within each stratum.

calculating FPC. However, using the sum of weights can sometimes result in a sample count being larger than the population count because the final weights are post-stratified to employment totals and not to establishment counts. Because of this, it was decided to use the sampling frame counts as population totals for calculating the FPC.

⁴ We also considered defining FPC as $FPC = \left(1 - \frac{M_r}{M}\right)$ where M_r is the sum of the measure of sizes (i.e. number of employees) of all responding establishments in the stratum and M is the sum of the measure of sizes of all responding and nonresponding eligible establishments. But considering the simplicity and for the sake of using the same FPC for HC and IC estimates, it was decided to use $FPC = \left(1 - \frac{n_r}{N}\right)$. However, if the sizes of certainties in the stratum do not vary substantially, the FPC is very similar in either case i.e., $\left(1 - \frac{M_r}{M}\right) \cong \left(1 - \frac{n_r}{N}\right)$.

If the parent agency includes sub-agencies (dependent government units), all of which are selected with certainty and the parent agency is treated as a PSU. The sampling stratum where the agency belongs is specified as the stratum. The number of parent agencies in each stratum is used as the population count (N) and the number of responding agencies is used as the sample count (n_r) in computing the FPC factor. Therefore, the FPC factor is calculated at the agency level, not at the sub-agency level, because all sub-agencies within a parent agency are selected with certainty.

For plan-level estimation, the same FPC factor is used with each agency specified as the cluster and the stratum for a plan is the same as the stratum for the agency. That is, the FPC factor is calculated based on the number of parent agencies i.e., both the numerator and the denominator in the FPC factor are the respective counts of parent agencies (not plans).

Note that the clustering of plans and sub-agencies was ignored in the previous method of variance estimation but is now accounted for in the new method of variance estimation which also incorporates the FPC.

4.2.2 Certainty Strata

For the government certainties, when a parent agency is selected all sub-agencies are also selected with certainty and nonresponse only happens at the sub-agency level, not at the agency level. Therefore, for government certainty strata, unlike the noncertainty strata, the FPC factor is defined at the sub-agency level with agency as a stratum and each sub-agency as a cluster. In other words, if a parent agency is selected with certainty and all sub-agencies are selected with certainty with the possibility of nonresponse only at the sub-agency level, then the parent agency is considered as a stratum and each sub-agency is considered as a PSU/cluster. Therefore, the total number of sub-agencies within an agency is used as the population count (N) and the number of responding sub-agencies is used as the sample count (n_r) in the calculation of the FPC factor.

For plan-level estimation, the FPC factor used for the sub-agencies are also used for the plans as plans are clustered within sub-agencies.

5.0 Impact of Incorporating FPC on MEPS-IC Variance

This section gives an indication of the effect of including FPC in the MEPS-IC variance estimation. The effect of FPC is assessed by sector (private versus government) within the certainty and noncertainty groupings. The effect is assessed separately for certainties and noncertainties because of differing directional impacts of the FPC on variances. The FPC is expected to have a decreasing effect on variance in noncertainty strata and an increasing effect in the certainty strata.

The comparison was made using 2015 MEPS-IC estimates in all published tables on the MEPS website that include both Privates and Governments tables; establishment and plan-based tables; national and state-level tables; and estimates of totals, ratios, and percentiles.

Also, note that this evaluation specifically focuses on the impact of the FPC along with the formation of variance strata for certainties to account for nonresponse (which allows for positive variances) but does not consider the additional impact of the clustering of plans or sub-agencies that is accounted for in the new method of variance computation.

5.1 Noncertainty Strata

For noncertainties, if the sampling rate is very high then the $FPC = (1 - n_r/N)$ is considerably lower than 1.0 and will notably reduce the variance. Therefore, we will analyze the distribution of realized sampling rates (n_r/N) for noncertainty strata within private and government sectors. If the sampling rates in strata that includes many establishments are high (say above 5%) then the FPC is likely to reduce the variances of the estimation cells where the strata with high sampling rates contribute.

Table 1 presents the distribution of realized sampling rates for all noncertainties in the private sector. The average sampling rate is about 0.6% and for about 90% of the establishments the sampling rate is about 1% or less. Only for about 1% of the establishments, the sample rate is 5% or more. So incorporating FPC in the variance estimation is not expected to make a noticeable impact in general in the private sector noncertainty strata.

Table 1. Distribution of Sampling Rate for all Noncertainties in the Private Sector in 2015 MEPS-IC

Weighted Moments			
N	41819	Sum Weights	7176007
Mean	0.006	Sum Observations	42277.693
Std Deviation	0.167	Variance	0.028
Skewness	3.397	Kurtosis	12.433
Uncorrected SS	1414.139	Corrected SS	1165.059
Coeff Variation	2833.113	Std Error Mean	0.000

Weighted Quantiles	
Level	Quantile
100% Max	0.500
99%	0.044
95%	0.018
90%	0.011
75% Q3	0.005
50% Median	0.003
25% Q1	0.002
10%	0.002
5%	0.002
1%	0.001
0% Min	0.001

Table 2 presents the distribution of sampling rates for all noncertainties in the government sector. The average sampling rate is 3.4% but for about 10% of the cases the sampling rate is greater than 8% while for about 5% of the cases the sample rate is greater than 13%. So the impact of FPC for the estimates derived from the government sector noncertainty establishments will be higher than for the noncertainty private sector establishments. For strata covering about 25% of all government agencies, the variance due to incorporating FPC would be reduced by at least 3.9%.

Table 2. Distribution of Sampling Rate for all Noncertainties in the Government Sector in 2015 MEPS-IC

Weighted Moments			
N	2310	Sum Weights	66568
Mean	0.034	Sum Observations	2248.703
Std Deviation	0.260	Variance	0.067
Skewness	1.084	Kurtosis	0.006
Uncorrected SS	231.806	Corrected SS	155.844
Coeff Variation	769.070	Std Error Mean	0.001

Weighted Quantiles	
Level	Quantile
100% Max	0.573
99%	0.239
95%	0.135
90%	0.086
75% Q3	0.039
50% Median	0.015
25% Q1	0.006
10%	0.004
5%	0.003
1%	0.002
0% Min	0.002

5.2 Certainty Strata

As mentioned before, previously the IC variance estimation procedure assigned a variance of zero for certainty strata. That means it implicitly included an FPC that was equal to zero since the sampling rate is 100% in certainty strata. What the previous procedure ignored is nonresponse that can considerably reduce the realized sampling rate since nonresponse is significant in many certainty strata. Although a nonresponse adjustment is applied to control the bias from nonresponse, the uncertainty due to the reduced sample size is not captured. The reduction in the sample size due to nonresponse can considerably change the assumed variance of zero in a certainty stratum to be substantially greater than zero. Now that the FPC will be calculated based on the realized sampling rate with nonresponse incorporated, the impact of FPC in certainty strata will depend on the nonresponse rate. Thus, for the certainty sector, to assess the impact of FPC we will check the distribution of the nonresponse rate or nonresponse propensity of the certainty cases on the frame. If the nonresponse rate in a certainty stratum is close to zero then the variance with FPC will be essentially zero. However, if the nonresponse rate is somewhat greater than zero then the variance will be greater than zero. Therefore, the variance in the certainty strata will be determined by the nonresponse rate. Also, note that the effect of FPC in the certainty strata is opposite to that in the noncertainty strata. In the certainty strata, incorporating the FPC will increase the variance from zero while in the noncertainty strata it will reduce the variance.

Table 3 presents the distribution of nonresponse rates of certainties in the private sector. It shows that nonresponse propensity is noticeably greater than zero for most cases. For 75% of the certainties in the private sector, the nonresponse propensity is greater than 30% and, for 95% of the certainties, the nonresponse propensity is greater than 10%. Therefore, the FPC will introduce a nonzero variance in almost all certainty estimation cells and in many cases it can be significantly greater than zero given that many certainty establishments are subject to high nonresponse propensities.

Table 3. Distribution of Nonresponse Rates of Certainties in the Private Sector in 2015 MEPS-IC

Weighted Moments			
N	137	Sum Weights	518.266
Mean	0.436	Sum Observations	226.216
Std Deviation	0.363	Variance	0.131
Skewness	0.059	Kurtosis	0.330
Uncorrected SS	116.622	Corrected SS	17.881
Coeff Variation	83.073	Std Error Mean	0.016

Weighted Quantiles	
Level	Quantile
100% Max	0.868
99%	0.868
95%	0.750
90%	0.631
75% Q3	0.589
50% Median	0.398
25% Q1	0.303
10%	0.221
5%	0.107
1%	0.000
0% Min	0.000

Table 4 shows the distribution of nonresponse propensities of certainty establishments in the government sector. For about 75% of the cases, the nonresponse propensities are greater than 14% and for about 50% of the cases, the nonresponse propensities are greater than 20%. So, the variance can be considerably greater than zero in many estimation cells.

Table 4. Distribution of Nonresponse Rates of Certainties in the Government Sector in 2015 MEPS-IC

Weighted Moments			
N	423	Sum Weights	886.406
Mean	0.196	Sum Observations	174.139
Std Deviation	0.185	Variance	0.034
Skewness	1.100	Kurtosis	5.669
Uncorrected SS	48.588	Corrected SS	14.378
Coeff Variation	93.956	Std Error Mean	0.006

Weighted Quantiles	
Level	Quantile
100% Max	0.667
99%	0.417
95%	0.417
90%	0.417
75% Q3	0.247
50% Median	0.203
25% Q1	0.148
10%	0.000
5%	0.000
1%	0.000
0% Min	0.000

6.0 Conclusion

The analysis indicates that incorporating FPC will have a noticeable impact on the variances in many strata of both certainties and noncertainties.

For noncertainty strata, since the sampling rates are generally low in most strata, the effect of incorporating FPC will be low overall. However, for many noncertainty strata, the realized sampling rates in both the private and government sectors are found to be non-negligible. About 10% of the establishments in the private sector were subject to a sampling rate of 1% or more, and about 10% of the government agencies were subject to a sampling rate of more than 8%, indicating that incorporating FPC will reduce the variance in some noncertainty strata.

For certainty strata, the effect of incorporating FPC will be more pronounced. Since the nonresponse propensity of most certainty cases are fairly high, the variance in almost all certainty strata will be considerably greater than the previously assumed value of zero. For the certainties in the private sector, 75% of establishments have a nonresponse propensity of 30% or more and 95% of establishments have a nonresponse propensity greater than 10%. For the certainties in the government sector, 75% of the cases have a nonresponse propensity of greater than 14%. Therefore, treating the certainty strata as noncertainty and incorporating the FPC based on realized sampling rate can capture the variance that was not historically accounted for.

However, since including an FPC will increase the variance in many certainty strata and will reduce the variance in many noncertainty strata, in estimation cells where both certainty and noncertainty establishments contribute, the impact of the FPC may cancel out to some extent. Moreover, the distribution of sampling rates or response rates analyzed above are at the establishment level but many different establishments from different strata (both certainty and noncertainty) contribute to an estimation cell. Hence, the impact of FPC may not be exactly as discussed above but this can be considered as a rough indication.

Also, unlike prior to the incorporation of FPC, since plans/government sub-agencies are now considered as clustered within establishments/parent agencies, it will have an increasing effect on the variances of relevant estimates. So again in some cases the variance reduction effect of incorporating the FPC will cancel out due to the accounting of clustering in the new variance estimation process.

A more accurate assessment of the impact of incorporating FPC can be made by producing variance estimates for estimation cells in some MEPS-IC tables with and without including FPC, which will be possible when the FPC is actually incorporated in the next production cycle i.e., for the 2016 estimates. Regardless of the magnitude of any impact, incorporating the FPC will produce more accurate and technically defensible standard errors for MEPS-IC estimates.

7.0 References

- Chowdhury, S. and Kashihara, D. (2017). A Comparison of Variance Estimates Using Random Group and Taylor Series Methods for a Large National Survey of Business. *Proceedings of the 2017 Joint Statistical Meetings Conference*, Baltimore, Maryland.
- Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.
- Deming, W.E. and Stephan, F.F. (1941). On the Interpretation of Censuses as Samples. *Journal of the American Statistical Association*, 36 (213), 45-49.
- Graubard B. and Korn E. (2002). Inference for Superpopulation Parameters Using Sample Surveys. *Statistical Science*, 17(1), 73-96.
- 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, Maryland.
http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr30/mr30.pdf
- Rust, K., Fuller, W., Stokes, L., and Kott, P. (2017). Finite Population Correction Factors (Panel Discussion). https://www.researchgate.net/publication/251618326_Finite_Population_Correction_Factors_Panel_Discussion.
- Sarndal, C., Swensson, B., and Wretman, J., (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag Inc.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag Inc.