Accounting for Multi-stage Sample Designs in Complex Sample Variance Estimation

Brady T. West, Michigan Program in Survey Methodology

Nationally representative samples of large populations often have complex design features for a variety of reasons (e.g., cost efficiency). For purposes of estimating sampling variances based on complex multi-stage sample designs involving stratification and cluster sampling, the sampling error codes provided by survey organizations in public use survey data files often assume "ultimate cluster selection" of individuals from primary sampling units (PSUs). "Ultimate clusters" are the ultimate aggregate samples of individual population elements that are selected from PSUs (Wolter, 2007, p. 33), possibly based on multiple stages of sample selection (e.g., counties as PSUs, area segments within counties, households within area segments, and individuals within households). The sample selection method that is often assumed for variance estimation in practice (possibly within strata) is a single-stage, with-replacement selection of ultimate clusters from all hypothetical ultimate clusters defined by a multi-stage design, where all units within the ultimate clusters are measured (i.e., there is no subsampling within the clusters, and the ultimate clusters are what are really being sampled in a single stage of selection).

This assumption is of course an approximation of the full multi-stage complex design, but turns out to be sufficient for estimating variances in practice. In fact, if PSUs were actually sampled with replacement at the first stage (again, possibly within strata) and arbitrary multi-stage selection schemes were used within the independently sampled PSUs, an unbiased estimator of the variance of an estimated sampled total only requires knowledge of the estimated totals for the sampled PSUs (or in other words, the estimated totals for the ultimate clusters); see Wolter (2007, Chapter 1). Furthermore, if the PSUs were actually sampled without replacement in a multi-stage design (which is typically the case in practice), standard weighted estimates of population totals are computed (using, for example, the Horvitz-Thompson estimator), and variances of these estimated totals are estimated under the assumption of a single-stage, with-replacement selection of ultimate clusters, the variance estimator only has a slight positive bias, making inferences conservative (see Kish, 1965, Section 5.3B). The failure to account for all stages of cluster sampling in complex multi-stage sample designs when performing design-based variance estimation often leads to confusion and debate among analysts, and this technical report aims to clarify why this technique is sufficient in practice.

For analysts to estimate variances in practice, at least two PSUs are needed within a first-stage sampling stratum for design-based variance estimation purposes. This is because unbiased variance estimators for estimated totals in multi-stage, stratified cluster samples are primarily driven by the variance of estimated cluster totals within a stratum. Some fairly common sample designs can therefore hinder variance estimation techniques commonly programmed in statistical software packages. For example, in multi-stage sample designs with one PSU selected per stratum (possibly for cost efficiency) and one ultimate cluster selected per PSU (possibly based on multiple stages of selection), there is

only one ultimate cluster available from each stratum. In addition, PSUs selected with certainty in a complex design should be treated as their own strata, leading again to strata with only a single PSU (and therefore only one ultimate cluster). The *random groups* method can be used to divide these ultimate clusters into *k* random groups (or *sampling error computation units*) for variance estimation purposes (Wolter, 2007, Chapter 2). These common design features lead to the need for a technique like the random groups method to create sampling error codes for variance estimation (e.g., Heeringa et al., 2010, Chapter 4).

Random group variance estimators also estimate variances as if PSUs were sampled with replacement, even if they may have been selected without replacement within first-stage sampling strata. The result is a slight positive bias in the variance estimator, which is sufficient for conservative variance estimation (Wolter, 2007, Chapter 2). In some cases, strata defined by certainty PSUs with similar features are *collapsed*, leading to pseudo-strata with ultimate clusters from two PSUs; this also introduces a slight bias in variance estimates. In practice, these positive biases in the variance estimators are slight, and lead to conservative inferences. In other cases, ultimate clusters from non-certainty PSUs might be *combined* into pseudo-strata for variance estimation, which does not bias variance estimates (but rather biases the *variance* of variance estimates).

Mathematically, variance estimators for sample totals based on complex sample designs featuring *without* replacement selection of PSUs at the first sampling stage are a function of finite population corrections (FPCs), which account for the proportion of a finite population that was not included in a sample selected without replacement, and *joint* probabilities of selection for sampled units (see Wolter, 2007, Chapter 1). Joint selection probabilities are often quite difficult to compute, meaning that survey agencies generally do not provide them to public users of the survey data. Larger sample sizes (relative to the size of a finite population) generally lead to FPCs that have the effect of reducing variance estimates (in the case of a without replacement design). Smaller sample sizes (again relative to the size of a finite population) generally result in minimal FPCs, which are usually ignored at all levels of a multi-stage design when sampling from large populations. Assumptions of *with-replacement* selection within primary stage strata also eliminate FPCs from variance estimators. Here is the key point: In a complex multi-stage sample involving stratification and without-replacement selection of PSUs, contributions to the variances of estimates of sample totals (the primary building blocks for common variance estimation techniques like Taylor Series Linearization) from later stages of cluster sampling are defined by a sum across first-stage strata of the (first-stage) stratum-specific sampling fractions multiplied by the additional variance contributions from the lower stages of selection (Canette, 2010; Cochran, 1977, p. 278-279). As a result, if the first-stage stratum-specific sampling fractions are small (leading to negligible finite population corrections), these contributions to variances from lower stages will be negligible and can be safely ignored. Furthermore, the (often positive) bias of the variance estimator assuming single-stage, with-replacement selection of ultimate clusters (see Wolter, 2007, Theorem 2.4.6) will be negligible as well.

These issues largely apply to Taylor Series Linearization, which is the default variance estimation technique for complex samples in many popular statistical software packages. Many survey data sets include a series of replicate weights that enable replicated variance estimation methods (jackknife repeated replication, balanced repeated replication, or bootstrapping), and do not include codes indicating strata or clusters at various levels of a complex design. These replicate weights are computed using first-stage sample design codes (or codes based on a sampling error computation model) to preserve respondent confidentiality and minimize risk of disclosure, and sampling error codes are generally not available in these data sets (or necessary) for theoretically appropriate replicated variance estimation methods (Wolter, 2007, Chapters 3-5).

Some software procedures designed for analysis of complex sample survey data (e.g., the SURVEY procedures in SAS) do not allow users to specify lower-stage sampling error codes for variance estimation, given that variance estimation based on first-stage sampling error codes for large complex samples has become so commonplace. Consider this statement from the SAS (V9.2) online technical documentation:

> The Taylor series linearization method is appropriate for all designs in which the first-stage sample is selected with replacement, or in which the first-stage sampling fraction is small, as it often is in practice. The Taylor series method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself. When there are clusters (PSUs) in the sample design, the procedures estimate the variance from the variation among the PSUs. When the design is stratified, the procedures pool stratum variance estimates to compute the overall variance estimate.
>
> For a multistage sample design, the Taylor series method uses only the first stage of the sample design. **Therefore, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling.**[1]

Other software procedures (e.g., the `svy` procedures in Stata, or the functions in the `survey` package in R) allow users to specify sample design information from multiple levels of a multi-stage sample design. Lumley (2010, Section 3.2.2) provides an example of how much larger variance estimates assuming single-stage, with-replacement selection of ultimate clusters can be when a multi-stage design featuring without-replacement selection of PSUs was actually used, and finds fairly negligible differences in standard errors. Although survey data producers often do not release these lower-stage design codes in order to preserve respondent confidentiality and limit the risk of statistical disclosure, this can lead to confusion among analysts regarding appropriate methods for variance estimation.

What follows is an example analysis of data from the first 10 quarters of Cycle 7 of the U.S. National Survey of Family Growth (NSFG), using the Stata software (which allows users to identify sample design codes at multiple levels of a multi-stage design). This example shows that the contribution of lower-stage clusters to variance estimates is at

[1] http://support.sas.com/rnd/app/da/new/dasurvey.html

best negligible in the NSFG, and does not need to be accounted for in practice by analysts.

Analysis Example Using Data from NSFG Cycle 7, Quarters 1-10

Briefly, the NSFG design involves selection of one PSU within each first-stage stratum, where PSUs are counties (or groups of counties) that have been classified into three major strata (large metro areas, other metro areas, and nonmetro areas). The 28 large metro areas are considered self-representing units, and are selected with certainty. The remaining PSUs are grouped by geography and population size into 80 first-stage strata, with one PSU selected from each of these strata. This design results in 108 first-stage strata, with one PSU selected from each. Within each selected PSU, there are several additional stages of selection (Lepkowski et al., 2010). This example focuses simply on the second stage of selection, or selection of area segments within PSUs, where area segments were census blocks (or groups of small census blocks). Housing units and respondents were eventually selected within blocks to define the "ultimate clusters" within each of the 108 selected PSUs.

Because only one ultimate cluster was selected within each PSU, area segments within each PSU are randomly grouped into either two or four SECUs for variance estimation purposes. Within the 28 self-representing PSUs, area segments are numbered systematically by sample selection order within sample domains as 1, 2, 3, 4, 1, 2, 3, … Area segments coded 1 and 3 are combined to form a pseudo-stratum with two SECUs (the '1' area segments and the '3' area segments). Area segments coded '2' and '4' are combined in a similar fashion to form a second pseudo-stratum with two SECUs, yielding two pseudo-strata with two SECUs each from the larger self-representing PSUs. Ultimate clusters in the smaller self-representing PSUs were simply divided into two SECUs by randomly grouping the area segments. This resulted in 36 pseudo-strata with 72 SECUs for variance estimation. For the remaining NSR PSUs, the 80 strata were inspected to identify groups of 4 PSUs that were as similar as possible. This led to 20 pseudo-strata with four SECUs each. The codes for these 152 SECUs and 56 pseudo-strata in total were scrambled by NSFG staff to define the sampling error codes in the NSFG data set for variance estimation purposes.

In most applications, these codes would be sufficient for appropriate design-based variance estimation. In this example, we also consider the area segments selected within each first-stage SECU as lower-level clusters, and account for this lower stage of selection in the variance estimation. This variable is not available in the NSFG public use data set for Quarters 1-10, given concerns about respondent confidentiality. We consider the following variables in this example, available for all sampled females in Quarters 1-10 (n = 7,356):

- Final Sampling Weight (FINALWGT30)
- First-Stage Sampling Error Stratum Codes (SEST)
- First-Stage Sampling Error Computation Unit Codes (SECU)
- Second-Stage Segment Codes within SECUs (SEGMENT)

- Ever been married (EVRMARRY)
- Ever had sex with a male (HADSEX)
- Currently using a birth control pill (PILL)

The following Stata `svyset` command processes these design variables for the subsequent `svy` analysis procedures. Consistent with the NSFG sample design, we add a negligible first-stage within-stratum sampling fraction (FPC1 = 0.0001) to the Stata code; a failure to do this would lead Stata to assume that units are sampled with replacement at the first stage, and all subsequent stages of selection specified in the `svyset` command would be ignored. We note that different software will handle this issue in different ways.

```
gen fpc1 = 0.0001

svyset secu [pweight=finalwgt30], strata(sest) fpc(fpc1) || segment
```

Note the locations in the syntax where the first-stage sampling error computation units (SECU) and the second-stage units (SEGMENT) would be identified. The second stage sampling units are identified after `||`, indicating a lower level of clustering (i.e., segments are nested within levels of SECU).

We now generate estimated proportions for the three NSFG variables, in addition to Linearized standard errors for the three weighted estimates:

```
svy: prop evrmarry hadsex pill
```

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      51        Number of obs    =      7356
Number of PSUs   =     124        Population size  =  61864498
                                  Design df        =        73

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
evrmarry     |
           0 |   .4527268   .0126088      .4275974    .4778561
           1 |   .5472732   .0126088      .5221439    .5724026
-------------+------------------------------------------------
hadsex       |
           1 |   .8605983   .0127321      .8352232    .8859734
           2 |   .1394017   .0127321      .1140266    .1647768
-------------+------------------------------------------------
pill         |
           1 |   .7277184   .0130585      .7016928    .7537441
           5 |   .2719962   .0130984      .2458912    .2981012
           8 |   .0002854   .0002052     -.0001236    .0006945
--------------------------------------------------------------
```

Next, we submit a `svyset` command to identify the first-stage sampling error codes only:

```
svyset secu [pweight=finalwgt30], strata(sest)
```

Note that no finite population correction or second-stage sampling unit has been specified in this command. We now once again generate estimated proportions and Linearized standard errors on the two survey variables:

```
svy: prop evrmarry hadsex pill
```

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      51        Number of obs    =      7356
Number of PSUs   =     124        Population size  =  61864498
                                  Design df        =        73

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
evrmarry     |
           0 |   .4527268   .0126088      .4275974    .4778561
           1 |   .5472732   .0126088      .5221439    .5724026
-------------+------------------------------------------------
hadsex       |
           1 |   .8605983   .0127323      .8352229    .8859737
           2 |   .1394017   .0127323      .1140263    .1647771
-------------+------------------------------------------------
pill         |
           1 |   .7277184   .0130587      .7016924    .7537444
           5 |   .2719962   .0130985      .2458908    .2981015
           8 |   .0002854   .0002052     -.0001236    .0006945
--------------------------------------------------------------
```

The estimated standard errors are virtually identical, indicating the minimal added contribution of this lower stage of cluster sampling to the overall design-based variance estimates in the NSFG.

Could this have been due to the fact that the assumed first-stage sampling fraction in the strata was too low? Consider the results when using FPC2 = 0.01:

```
gen fpc2 = 0.01

svyset secu [pweight=finalwgt30], strata(sest) fpc(fpc2) || segment

svy: prop evrmarry hadsex pill
```

```
(running proportion on estimation sample)
```

```
Survey: Proportion estimation

Number of strata =      51        Number of obs    =      7356
Number of PSUs   =     124        Population size  =  61864498
                                  Design df        =        73

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
evrmarry     |
           0 |   .4527268   .0126059      .4276033    .4778502
           1 |   .5472732   .0126059      .5221498    .5723967
-------------+------------------------------------------------
hadsex       |
           1 |   .8605983   .0127167      .8352539    .8859428
           2 |   .1394017   .0127167      .1140572    .1647461
-------------+------------------------------------------------
pill         |
           1 |   .7277184   .0130406      .7017286    .7537082
           5 |   .2719962   .0130801      .2459276    .2980647
           8 |   .0002854   .0002052     -.0001236    .0006944
--------------------------------------------------------------
```

Again, we see negligible differences. In fact, the estimated standard errors based on the first-stage sampling error codes only are *larger* than the estimated standard errors when taking into account the subsequent stage of sample selection (which is not surprising, due to the expected positive bias in the variance estimator assuming single-stage, with replacement selection of ultimate clusters); this illustrates how the with-replacement assumption will typically lead to (slightly) conservative inferences. What about adding a finite population correction of 0.01 for sampling at the *second* stage?

```
svyset secu [pweight=finalwgt30], strata(sest) fpc(fpc2) ///
      || segment, fpc(fpc2)

svy: prop evrmarry hadsex pill
```

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      51        Number of obs    =      7356
Number of PSUs   =     124        Population size  =  61864498
                                  Design df        =        73

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
evrmarry     |
           0 |   .4527268   .0126053      .4276045     .477849
           1 |   .5472732   .0126053       .522151    .5723955
-------------+------------------------------------------------
hadsex       |
```

```
             1 |   .8605983   .0127163      .8352549    .8859418
             2 |   .1394017   .0127163      .1140582    .1647451
-------------+------------------------------------------------
pill         |
             1 |   .7277184   .0130401      .7017295    .7537073
             5 |   .2719962   .0130796      .2459286    .2980638
             8 |   .0002854   .0002052     -.0001235    .0006944
------------------------------------------------------------------
```

We once again see negligible differences.

Conclusions

There is no guarantee that the results illustrated here will hold for data sets from all large surveys. Design-based methods for variance estimation based on ultimate cluster sampling within strata have a long history of empirical study and application in many scientific fields, and are "tried and true" methods for correctly accounting for complex sampling designs when estimating sampling variances. These methods are also easy for analysts in a variety of fields to implement. More sophisticated analysts concerned about the contributions of cluster sampling at lower stages of a multi-stage design to variance estimates are urged to consider the following points:

1. Is the survey organization that produced the data set willing to provide codes for the lower-stage sampling units, or will this present a disclosure risk?
2. Is the survey organization that produced the data set willing to provide finite population corrections for strata at various levels of the multi-stage design?
3. Is software available that correctly accounts for the lower-stage sampling units and finite population corrections (e.g., the `svy` commands in Stata)?

Given this information, analysts can use the methods described in the example above to examine the sensitivity of their inferences to assumptions about negligible contributions of lower-stage cluster sampling.

Finally, we note that multilevel modeling methods enable analysts to take a more "model-based" approach to the analysis of a complex sample survey data set, accounting for multiple levels of cluster sampling. Analysts interested in decomposing overall variance estimates into components of variance due to different levels of sample selection (rather than simply estimating total sample variances of estimates) will find multilevel models useful for this type of application. Interested readers can refer to Asparouhov and Muthen (2006), Carle (2009), or Rabe-Hesketh and Skrondal (2006) for more details on these types of approaches.

References

Asparouhov, T. & Muthen, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meetings in Seattle, August 2006*. ASA section on Survey Research Methods, 2718-2726.

Canette, I. (2010). Analysis of complex survey data in Stata. *Paper presented at the 2010 Mexican Stata Users Group Meeting, April 29, 2010.*

Carle, A.C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 1471-2288-9-49.

Cochran, W.G. (1977). *Sampling Techniques*. Wiley.

Heeringa, S.G., West, B.T. and Berglund, P.A. (2010). *Applied Survey Data Analysis*. Chapman and Hall / CRC Press.

Kish, L. (1965). *Survey Sampling*. Wiley.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M., and Van Hoewyk, J. (2010). The 2006-2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey. National Center for Health Statistics, Vital and Health Statistics, 2(150), June 2010.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* Wiley.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society-Series A*, 169, 805-827.

Wolter, K.M. (2007). *Introduction to Variance Estimation, Second Edition*. Springer-Verlag.