Design-corrected Variance Estimation of NSECE Statistics (revised 2/25/16)

The NSECE is a probability-based study that relies on sampling methods to represent the target population. This means that a representative sample was randomly selected in such a way that is possible to make inferences about the population of interest, provided users make use of survey design variables. In the NSECE study, some characteristics of the population are of particular interest (for example, low income areas); as a consequence, some subgroups of the population were disproportionally selected.

Additionally, survey response rates were not uniform across different sample units. In order to obtain valid inferences for statistics of interest (e.g., percentages, means, totals, ratios, regression coefficients), cases need to be weighted by the inverse of their probability of selection and response. Statistics derived from the NSECE study without proper weighting adjustment will result in biased estimates. The following section offers programming examples on how to calculate two common statistics —totals and percentages— accounting for survey design variables.

The data used for examples come from the Inter-University Consortium for Political and Social Research at University of Michigan, as documented in study ICPSR Study No. 35519 "Center-based Public-Use File." The numbers reported as part of this reference guide use Exhibit 1 of "Characteristics of Center-based Early Care and Education Programs: Initial Findings from the National Survey of Early Care and Education (NSECE); OPRE Report #2014-73a.[1] Results from this guide do not exactly match those reported in OPRE Report #2014-73a. This is because Exhibit 1 figures were calculated with data prior to top-coding procedures that were implemented to protect against disclosure. Nonetheless, the numbers produced examples are close enough to give the user a contextual reference. A summary table with results from the programming examples is offered at the end of Example 1 and 2. This summary table follows the same structure as Exhibit 1 OPRE Report #2014-73a.

Three pieces of information are needed in the variance estimation process: weights, strata, and primary sampling units (PSU), also referred to as clusters. Variables proxying for these survey-design features are included as part of the NSECE datasets, as documented in Table A below. [Specifically, the variables VSTRAT and VPSU constructed for variance estimation do not indicate geographic units such as counties or provider clusters as those terms are used in the NSECE sample design.] These survey design variables need to be used in combination, to obtain valid inferences from NSECE data. Examples 1 and 2 provide an illustration of the use of these survey design variables.

[1] http://www.acf.hhs.gov/sites/default/files/opre/characteristics_of_cb_ece_programs_111014.pdf

Although these examples use the Center-based Provider Data, the same procedures would apply to estimates using any of the NSECE data files.

Table A. Survey-Design Variable Names for Four NSECE Data Files

Survey Design Variables	Households and Children[2]	Home-Based Providers (Informal and Formal)	Center-Based Providers[3]	Workforce
Weight	HH_METH_ WEIGHT	HB_METH_WEIGHT	CB_METH_WEIGHT	WF_METH_WEIGHT
Stratum	HH_METH_VSTRATUM	HB_METH_VSTRATUM	CB_METH_VSTRATUM	WF_METH_VSTRATUM
PSU	HH_METH_VPSU	HB_METH_VPSU	CB_METH_VPSU	WF_METH_VPSU

Example 1. Total Number of Children Enrolled by Single Age Category

This example provides users with variables used to construct Exhibit 1 of "Characteristics of Center-based Early Care and Education Programs: Initial Findings from the National Survey of Early Care and Education (NSECE); OPRE Report #2014-73a. This section shows how to calculate total number of children enrolled in center-based programs that provide care and education for children age birth through five years and not yet in kindergarten (Exhibit 1, first column). The dataset utilized is 35519-0008-Data.dta. It was downloaded in its original Stata format and analyzed in Stata version 14. No formatting or additional data management manipulations such as recoding are included in the examples, except when necessary and as annotated. The example shows how to calculate the count of enrolled children by single year age of child (that is, less than one year old, one year old, two year old, three, four, and 5 year old not yet in kindergarten).

Stata

In Stata, the set-up stage for survey design features is separated from the estimation stage. Once the user has successfully opened up any dataset, the command **`svyset`** is used to declare the weight, stratum and PSU variable. After declaring relevant design variables in this step (**`svyset`**), the variance estimation process of various statistics should be conducted using the command **`svy:`** prefix.

[2] The HH_ WEIGHT variable is for household-level analyses. The HHC_METH_WEIGHT_X (X=1-9) variables can be used for child-level analyses in the HH-file. HHC_METH_WEIGHT is the name of the child-level weight in the Child-level QT file.

3 See documentation for discussion of some variables not appropriately used with this weight. That is, classroom level analysis should be performed using CB_WEIGHT_CLSM.

Stata typically allows the user to apply survey commands similar to non-survey commands after the **svy:** prefix. This guide provides examples of typical analyses[4] using variables designed for weights (**CB_METH_WEIGHT**) strata (**CB_METH_VSTRATUM**) and clusters (**CB_METH_VPSU**), respectively.

Step 1: Setup of Survey-Design Variables

Stata command	Annotation
`cd "C:\MyDir"`	←Set up directory
`use "35519-0008-Data.dta", clear`	←Load data into Stata
`svyset [pweight=CB_METH_WEIGHT], strata(CB_METH_VSTRATUM) psu(CB_METH_VPSU) singleunit(scaled)`	←Set up survey design variables (This step is usually declared once at the beginning of the session)
`svydes`	← This command displays a table of strata and sampling units in PSUs. This command is optional

Step 2: Calculate Point Estimate and Standard Error for Subgroup of Interest
(Total number of children enrolled in center-based programs that provide care and education for children age birth through five years and not yet in kindergarten)

Stata command	Annotation
`svy, subpop(if CB_AGECAT_SERVE_0TO5==1 & CB_AGECAT_SERVE_0==1 & CB_AGECAT_TOTENROLL_TC_0>=0):total CB_AGECAT_TOTENROLL_TC_0`	←Calculate count of children in subgroup "less than one year old." This number correspond to first column, first row, in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Total estimation

Number of strata =      81        Number of obs    =      8,265
Number of PSUs   =   2,771        Population size  =    141,709
                                  Subpop. no. obs  =      2,926
                                  Subpop. size     = 46,564.563
                                  Design df        =      2,690

------------------------------------------------------------------------------
                             |             Linearized
                             |      Total   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
CB_AGECAT_TOTENROLL_TC_0     |   405968.8   29441.81       348238    463699.7
------------------------------------------------------------------------------
```

[4] The reader may consult additional survey commands at http://www.stata.com/capabilities/survey-commands/

svy, subpop(if CB_AGECAT_SERVE_0TO5==1 & CB_AGECAT_SERVE_1==1 & CB_AGECAT_TOTENROLL_TC_1>=0):total CB_AGECAT_TOTENROLL_TC_1

←Calculate count of children in subgroup "One year old."
This number correspond to first column, second row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Total estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Subpop. no. obs =      3,568
                                  Subpop. size    = 55,892.723
                                  Design df       =      2,690

------------------------------------------------------------------------------
                             |             Linearized
                             |      Total   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
CB_AGECAT_TOTENROLL_TC_1     |   585489.5   38028.26      510921.9    660057.1
------------------------------------------------------------------------------
```

svy, subpop(if CB_AGECAT_SERVE_0TO5==1 & CB_AGECAT_SERVE_2==1 & CB_AGECAT_TOTENROLL_TC_2>=0):total CB_AGECAT_TOTENROLL_TC_2

←Calculate count of children in subgroup "Two year old."
This number correspond to first column, third row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Total estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Subpop. no. obs =      4,612
                                  Subpop. size    = 67,915.233
                                  Design df       =      2,690

------------------------------------------------------------------------------
                             |             Linearized
                             |      Total   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
CB_AGECAT_TOTENROLL_TC_2     |   903713.4   53868.74        798085     1009342
------------------------------------------------------------------------------
```

Stata command	Annotation
svy, subpop(if CB_AGECAT_SERVE_0TO5==1 & CB_AGECAT_SERVE_3==1 & CB_AGECAT_TOTENROLL_TC_3>=0):total CB_AGECAT_TOTENROLL_TC_3	←Calculate count of children in subgroup "Three year old." This number correspond to first column, fourth row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Total estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Subpop. no. obs =      6,742
                                  Subpop. size    = 108,449.86
                                  Design df       =      2,690
```

	Linearized			
	Total	Std. Err.	[95% Conf.	Interval]
CB_AGECAT_TOTENROLL_TC_3	2110053	95656.36	1922485	2297620

svy, subpop(if CB_AGECAT_SERVE_0TO5==1 & CB_AGECAT_SERVE_4==1 & CB_AGECAT_TOTENROLL_TC_4>=0):total CB_AGECAT_TOTENROLL_TC_4

←Calculate count of children in subgroup "Four year old."
This number correspond to first column, fifth row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Total estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Subpop. no. obs =      6,883
                                  Subpop. size    = 114,335.05
                                  Design df       =      2,690
```

	Linearized			
	Total	Std. Err.	[95% Conf.	Interval]
CB_AGECAT_TOTENROLL_TC_4	2241592	98702.64	2048052	2435133

svy, subpop(if CB_AGECAT_SERVE_0TO5==1 & CB_AGECAT_SERVE_5==1 & CB_AGECAT_TOTENROLL_TC_5>=0):total CB_AGECAT_TOTENROLL_TC_5

←Calculate count of children in subgroup "Five year old (not yet in Kindergarten)."
This number correspond to first column, sixth row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Total estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Subpop. no. obs =      2,939
                                  Subpop. size    = 49,096.171
                                  Design df       =      2,690
```

	Linearized			
	Total	Std. Err.	[95% Conf.	Interval]
CB_AGECAT_TOTENROLL_TC_5	461026.9	34629.87	393123.1	528930.8

Example 2. Percent of Programs by Single Age Category

Similar to Example 1 above, this example provides users with variables used to construct Exhibit 1 of "Characteristics of Center-based Early Care and Education Programs: Initial Findings from the National Survey of Early Care and Education (NSECE); OPRE Report #2014-73a.[5] This example demonstrates how to calculate the percent of programs that provide care and education for children age birth through five years and not yet in kindergarten (Exhibit 1, last column). The dataset utilized is the same as in Example 1 (35519-0008-Data.dta) and also analyzed in Stata version 14. No formatting or additional data management manipulations such as recoding are included in these examples, except when necessary and as annotated. The example shows how to obtain the percent of programs with enrolled children by single age of child (that is, less than one year old, one year old, two year old, three, four, and 5 year old not year in kindergarten). The programming solution computes such statistics through Stata's command **`ratio`**.

Step 1: Setup of Survey-Design Variables	
(Step 1 can be skipped if survey design variables have been previously declared in Example 1, and are in use for current Stata session)	
Stata command	Annotation
`cd "C:\MyDir"`	←Set up directory
`use "35519-0008-Data.dta", clear`	←Load data into Stata
`svyset [pweight=CB_METH_WEIGHT], strata(CB_METH_VSTRATUM) psu(CB_METH_VPSU) singleunit(scaled)`	←Set up survey design variables (This step is usually declared once at the beginning of the session)
`svydes`	← This command displays a table of strata and sampling units in PSUs. This command is optional

[5] http://www.acf.hhs.gov/sites/default/files/opre/characteristics_of_cb_ece_programs_111014.pdf

Step 2: Calculate Point Estimate and Standard Error for Subgroup of Interest
(Percent of programs that provide care and education for children age birth through five years and not yet in kindergarten)

Stata command

```
gen n=1

gen Program_0=0
replace Program_0=1 if
CB_AGECAT_SERVE_0TO5==1 &
CB_AGECAT_SERVE_0==1 &
CB_AGECAT_TOTENROLL_TC_0>=0

svy: ratio Program_0 /n
```

Annotation

←Create variable to set programs as individual units. Also, create variable to indicate appropriate center-based program; that is, serving children in subgroup "less than one year old."
Calculate percentage of interest as a ratio (Results in Table B of this document were multiplied by 100 to be displayed as percentage). This number corresponds to the last column, first row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Ratio estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Design df       =      2,690

     _ratio_1: Program_0/n
```

	Ratio	Linearized Std. Err.	[95% Conf.	Interval]
_ratio_1	.3285928	.0144439	.3002705	.3569152

```
gen Program_1=0
replace Program_1=1 if
CB_AGECAT_SERVE_0TO5==1 &
CB_AGECAT_SERVE_1==1 &
CB_AGECAT_TOTENROLL_TC_1>=0

svy: ratio Program_1 /n
```

←Create variable to indicate appropriate center-based program; that is, serving children in subgroup "one year old."
Calculate percentage of interest as a ratio.
This number corresponds to the last column, second row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Ratio estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Design df       =      2,690

     _ratio_1: Program_1/n
```

	Ratio	Linearized Std. Err.	[95% Conf.	Interval]
_ratio_1	.394419	.0149799	.3650457	.4237923

Stata command	Annotation
gen Program_2=0 **replace Program_2=1 if** **CB_AGECAT_SERVE_0TO5==1 &** **CB_AGECAT_SERVE_2==1 &** **CB_AGECAT_TOTENROLL_TC_2>=0** **svy: ratio Program_2 /n**	←Create variable to set programs as individual units. Also, create variable to indicate appropriate center-based program; that is, serving children in subgroup "two year old." Calculate percentage of interest as a ratio. This number corresponds to the last column, third row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Ratio estimation

Number of strata =      81        Number of obs    =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Design df       =      2,690

     _ratio_1: Program_2/n

------------------------------------------------------------------------------
             |                Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _ratio_1 |   .4792584   .0156841      .4485043    .5100126
------------------------------------------------------------------------------
```

gen Program_3=0 **replace Program_3=1 if** **CB_AGECAT_SERVE_0TO5==1 &** **CB_AGECAT_SERVE_3==1 &** **CB_AGECAT_TOTENROLL_TC_3>=0** **svy: ratio Program_3/n**	←Create variable to set programs as individual units. Also, create variable to indicate appropriate center-based program; that is, serving children in subgroup "three year old." Calculate percentage of interest as a ratio. This number corresponds to the last column, fourth row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Ratio estimation

Number of strata =      81        Number of obs    =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Design df       =      2,690

     _ratio_1: Program_3/n

------------------------------------------------------------------------------
             |                Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _ratio_1 |   .7652997   .0139978      .7378522    .7927473
------------------------------------------------------------------------------
```

Stata command	Annotation
gen Program_4=0 **replace Program_4=1 if** **CB_AGECAT_SERVE_0TO5==1 &** **CB_AGECAT_SERVE_4==1 &** **CB_AGECAT_TOTENROLL_TC_4>=0** **svy: ratio Program_4/n**	←Create variable to set programs as individual units. Also, create variable to indicate appropriate center-based program; that is, serving children in subgroup "four year old." Calculate percentage of interest as a ratio. This number corresponds to the last column, fifth row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Ratio estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Design df       =      2,690

     _ratio_1: Program_4/n

------------------------------------------------------------------------------
             |                Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _ratio_1 |   .8068298    .012138      .7830291    .8306306
------------------------------------------------------------------------------
```

gen Program_5=0 **replace Program_5=1 if** **CB_AGECAT_SERVE_0TO5==1 &** **CB_AGECAT_SERVE_5==1 &** **CB_AGECAT_TOTENROLL_TC_5>=0** **svy: ratio Program_5/n**	←Create variable to set programs as individual units. Also, create variable to indicate appropriate center-based program; that is, serving children in subgroup "Five year old (not yet in Kindergarten)." Calculate percentage of interest as a ratio. This number corresponds to the last column, sixth row in Exhibit 1 OPRE Report #2014-73a.

Result:

```
Survey: Ratio estimation

Number of strata =      81        Number of obs   =      8,265
Number of PSUs   =   2,771        Population size =    141,709
                                  Design df       =      2,690

     _ratio_1: Program_5/n

------------------------------------------------------------------------------
             |                Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _ratio_1 |   .3464577   .0129736      .3210184    .3718969
------------------------------------------------------------------------------
```

Table B summarizes results obtained through these Example 1 and 2 based on variables used in Exhibit 1 OPRE Report #2014-73a.

Table B. Summary of descriptive statistics. Programs and Children Enrolled by Single-year Age of Child

	Children Enrolled			Programs		
	Count	Standard Error	Percent of Enrolled Children Birth through 5 Years (Not included in example)	Count of Programs Serving Age Group (Not included in example)	Standard Error (Not included in example)	Percent of Programs Serving Age Group
<1 year old	405,968	29,441	---	---	---	32.8
1 year old	585,489	38,028	---	---	---	39.4
2 year old	903,713	53,868	---	---	---	47.9
3 year old	2,110,053	95,656	---	---	---	76.5
4 year old	2,241,592	98,702	---	---	---	80.6
5 year old (not yet in kindergarten)	461,026	34,629	---	---	---	34.6
Total children birth through 5 years (not yet in kindergarten)	---	---	---	---	---	---
School-age (including kindergarten)	---	---	---	---	---	---