Example 2: Variance estimates for Means using SAS (9.4) and STATA (14)

Mean Number of Children Ever Born, by Place of Residence for Women 20-49 Years of Age

Following are SAS and STATA programs and output for an analysis of the mean number of children born to women 20-49 years of age in the 2015-2017 NSFG female respondent file, by place of residence.

The estimates and standard errors are equivalent across SAS and STATA.

In these programs, variables in uppercase represent variables as named on the data files. Variables in lowercase represent variables that were created as part of this program. Library and file names are generic; the user must apply names specific to his/her computing environment. Formatting and library options are not presented since preferences will vary across user organizations. SAS format statements could be used instead of creating new variables for some examples shown here.

SAS 9.4

The DATA and SET steps create a dataset for females that contains the variables to be used in the analysis and the subpopulation indicator variable (agepop) that is used to identify women ages 20-49 years of age. When producing estimates for population subgroups (such as women ages 20-49 as shown here), it is important to read in the entire data set first. An indicator, or subpopulation, variable (like agepop used here) should be created to identify your subgroup of interest within your survey procedure. If the data are subset without first reading in the entire data set, then empty clusters may be lost, and you may have error messages when running your program and incorrect variance estimates. It is a good idea to verify the number of clusters and strata in your output to be sure you are reading the entire data set.

The PROC SURVEYMEANS step produces a table of weighted means for the variable specified in the VAR statement (PARITY) by place of residence (METRO) by using the DOMAIN statement. The WEIGHT statement identifies the weight variable (WGT2015_2017) to be used in estimating the means. PROC SURVEYMEANS calculates standard errors appropriate to the complex sample design variables specified in the STRATUM and CLUSTER statements. The NOMCAR option is included in this PROC SURVEYMEANS example even though there are no missing values. SAS documentation can provide more information about the NOMCAR option.

SAS Program

```
data NSFG.EX2;
set NSFG.FEMALES;

**Create a variable for your subpopulation of ages 20-49;
agepop=0;
if ager ge 20 then agepop=1;
run;

proc surveymeans nomcar;
stratum sest;
cluster secu;
domain agepop*metro;
var parity;
weight WGT2015_2017;
run;
```

SAS Output

```
Mean Numbers of Children Ever Born (PARITY) by Place of Residence for Women Ages 20-49

The SURVEYMEANS Procedure

            Data Summary

Number of Strata                  18
Number of Clusters                72
Number of Observations          5554
Sum of Weights              72218086


      Variance Estimation

Method           Taylor Series
Missing Values          NOMCAR

                                                   Statistics

                                                                              Std Error
Variable    Label                                     N           Mean         of Mean       95% CL for Mean
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
PARITY      TOTAL NUMBER OF LIVE BIRTHS            5554        1.280819        0.033136    1.21438593 1.34725258
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

The SURVEYMEANS Procedure

                                            Domain Statistics in agepop*METRO

          PLACE OF RESIDENCE
          (METROPOLITAN-                                                                                         Std Error
agepop    NONMETROPOLITAN)         Variable    Label                                     N            Mean         of Mean
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
yes       Principal city of MSA    PARITY      TOTAL NUMBER OF LIVE BIRTHS            1880        1.444937        0.074527
          Other MSA                PARITY      TOTAL NUMBER OF LIVE BIRTHS            2075        1.458681        0.065210
          Not MSA                  PARITY      TOTAL NUMBER OF LIVE BIRTHS             675        1.549797        0.079701
no        Principal city of MSA    PARITY      TOTAL NUMBER OF LIVE BIRTHS             317        0.052007        0.014994
          Other MSA                PARITY      TOTAL NUMBER OF LIVE BIRTHS             479        0.025674        0.011168
          Not MSA                  PARITY      TOTAL NUMBER OF LIVE BIRTHS             128        0.057442        0.027233
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

```
                                            Domain Statistics in agepop*METRO

          PLACE OF RESIDENCE
          (METROPOLITAN-
agepop    NONMETROPOLITAN)         Variable       95% CL for Mean
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
yes       Principal city of MSA    PARITY      1.29551944 1.59435436
          Other MSA                PARITY      1.32794190 1.58941924
          Not MSA                  PARITY      1.39000536 1.70958897
no        Principal city of MSA    PARITY      0.02194546 0.08206846
          Other MSA                PARITY      0.00328332 0.04806389
          Not MSA                  PARITY      0.00284319 0.11203987
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

STATA 14.0

STATA Program

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (WGT2015_2017), strata (SEST), and cluster (SECU) variables to be used in by STATA in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command.

The *generate* and *replace* statements create the variable indicating the subpopulation of women ages 20 and older. When producing estimates for population subgroups (such as women ages 20 and older as shown here), it is important to read in the entire data set first. An indicator, or subpopulation, variable (like agepop used here) should be created to identify your subgroup of interest within your survey procedure. If the data are subset without first reading in the entire data set, then empty clusters may be lost, and you may have errors in your program and incorrect estimates. It is a good idea to verify the number of clusters and strata in your output to be sure you are reading the entire data set.

The *svy: mean* command produces estimated weighted means for each of the levels of the by variable METRO to show means separately by place of residence by using the over statement. As with most programming, there are multiple options to get the results you need. For example, STATA also has the option to use a *subpop* command within *svy: mean* (svy, subpop(varname): mean varname). The estimates provided are appropriate to the complex sample design identified by the *svyset* command.

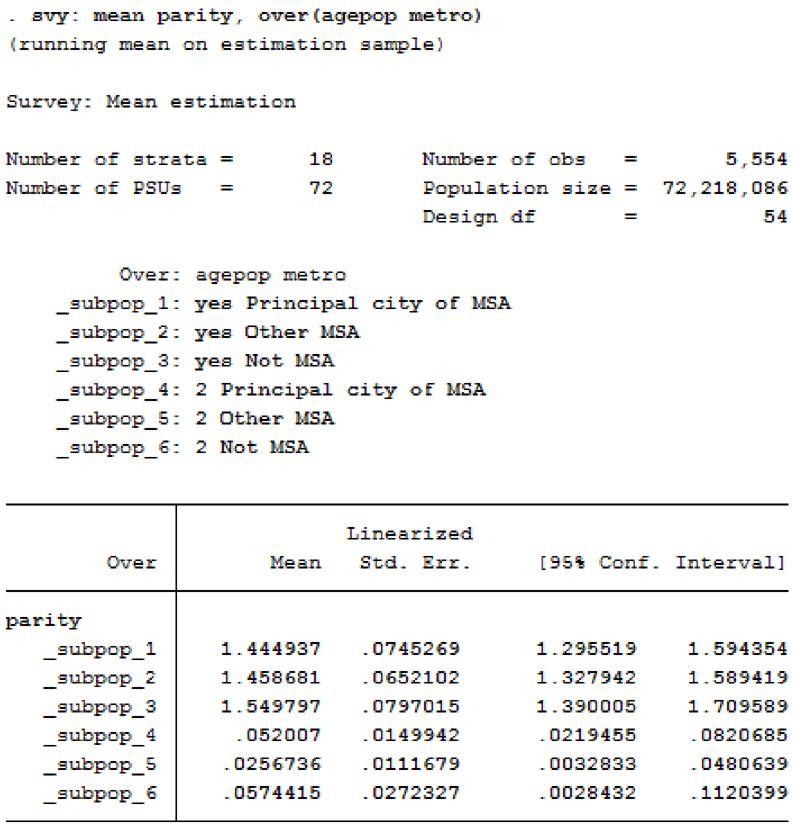
```
use “EX2.DTA”

svyset [pweight=WGT2015_2017], strata(SEST) psu(SECU)

* create a variable for your subpopulation of ages 20 and older
generate agepop=0
replace agepop=1 if AGER>=20

svy: mean parity, over(agepop metro)
```

STATA Output



```
. svy: mean parity, over(agepop metro)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      18       Number of obs   =      5,554
Number of PSUs   =      72       Population size = 72,218,086
                                 Design df       =         54

        Over: agepop metro
    _subpop_1: yes Principal city of MSA
    _subpop_2: yes Other MSA
    _subpop_3: yes Not MSA
    _subpop_4: 2 Principal city of MSA
    _subpop_5: 2 Other MSA
    _subpop_6: 2 Not MSA

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
parity       |
   _subpop_1 |   1.444937   .0745269      1.295519    1.594354
   _subpop_2 |   1.458681   .0652102      1.327942    1.589419
   _subpop_3 |   1.549797   .0797015      1.390005    1.709589
   _subpop_4 |    .052007   .0149942      .0219455    .0820685
   _subpop_5 |   .0256736   .0111679      .0032833    .0480639
   _subpop_6 |   .0574415   .0272327      .0028432    .1120399
--------------------------------------------------------------
```