Example 3: Variance Estimates for Percentages using SAS (9.4) and STATA (14)

Percentage of Men 20-49 Years of Age Who Have Ever Had One or More Biological Children, by Hispanic Origin and Race

Following are SAS and STATA programs and output for an analysis of the percentage of men aged 20-49 in the 2017-2019 NSFG male file who have ever fathered one or more biological children, tabulated by Hispanic origin and race.

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 containing variables from the male dataset to create a binary variable indicating whether the respondent fathered one or more biological children (biokidsx) based on the computed variable BIOKIDS. A subpopulation indicator for men ages 20-49 is also created. When producing estimates for population subgroups (such as men 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 estimation of variance. 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 SURVEYFREQ step produces a cross-tabulation of unweighted and weighted cell counts for the variables HISPRACE2 and biokidsx specified in the TABLE statement. The WEIGHT statement identifies the weight variable WGT2017_2019. PROC SURVEYFREQ calculates standard errors appropriate to the complex sample design specified by the STRATUM and CLUSTER statements. The specification of ROW in the TABLE statement limits the cell counts and percentages to the row. The NOMCAR option is included in this PROC SURVEYFREQ example even though there are no missing values on variables in the TABLE statement. Data users should consult official SAS documentation for more information about the NOMCAR option and options in the TABLE statement.

SAS Program

```
data EX3;
set NSFG.MALES (keep=CASEID BIOKIDS AGER HISPRACE2 SEST SECU WGT2017_2019);

if BIOKIDS gt 0 then biokidsx=1;
else biokidsx=0;

**create a variable for subpopulation of ages 20 and older;
agepop=0;
if AGER ge 20 then agepop=1;
run;

proc surveyfreq nomcar;
stratum SEST;
cluster SECU;
table agepop*HISPRACE2*biokidsx / ROW NOCELLPERCENT nosparse;
weight WGT2017_2019;
run;
```

SAS Output (output not shown for subpopulation variable agepop=no)

```
NSFG 2017-2019 Percentage of Males 20-49 Who Have Ever Fathered One or More Children by Hispanic Origin and Race


The SURVEYFREQ Procedure

            Data Summary

Number of Strata                  18
Number of Clusters                72
Number of Observations          5206
Sum of Weights              72221885


      Variance Estimation

Method            Taylor Series
Missing Values           NOMCAR

                                             Table of HISPRACE2 by biokidsx
                                              Controlling for agepop=yes

                                                                      Weighted    Std Err of         Row     Std Err of
HISPRACE2                              biokidsx        Frequency     Frequency      Wgt Freq     Percent    Row Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Hispanic                               none                  453       5545037        583939     41.7510         2.4097
                                       one or more           527       7736166        926830     58.2490         2.4097

                                       Total                 980      13281203       1367586    100.0000
-----------------------------------------------------------------------------------------------------------------------
Non-Hispanic White, Single Race        none                 1117      17595845       1800755     50.0512         2.5350
                                       one or more           937      17559813       1464061     49.9488         2.5350

                                       Total                2054      35155658       2755586    100.0000
-----------------------------------------------------------------------------------------------------------------------
Non-Hispanic Black, Single Race        none                  346       3347154        340450     45.9811         3.2261
                                       one or more           371       3932261        538484     54.0189         3.2261

                                       Total                 717       7279415        752872    100.0000
-----------------------------------------------------------------------------------------------------------------------
Non-Hispanic Other or Multiple Race    none                  237       3855866        665146     58.2917         4.5976
                                       one or more           186       2758911        351470     41.7083         4.5976

                                       Total                 423       6614777        821422    100.0000
-----------------------------------------------------------------------------------------------------------------------
Total                                  none                 2153      30343903       2426240
                                       one or more          2021      31987150       1849464

                                       Total                4174      62331053       3434390
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

STATA 14

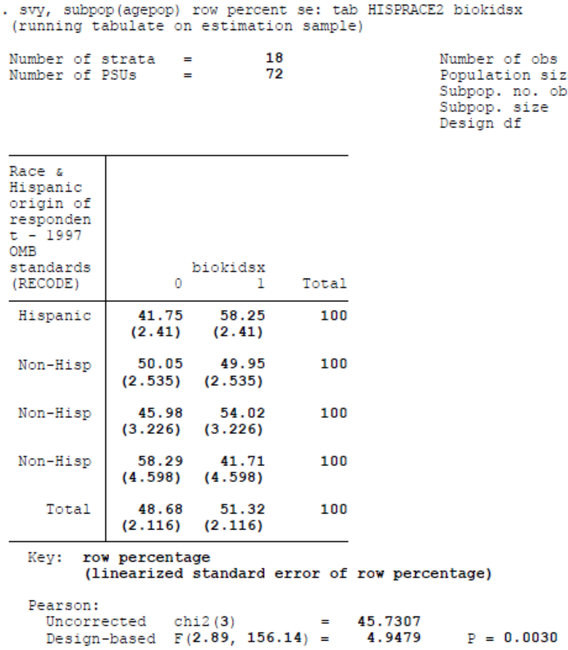
The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (WGT2017_2019), strata (SEST), and cluster (SECU) variables to be used in STATA in estimation. These settings are saved for the current session but can be cleared by entering the *clear* command. The *generate* and *replace* statements create the variable biokidsx, a binary indicator of whether the respondent fathered one or more biological children (biokidsx) based on the computed variable BIOKIDS. A subpopulation indicator for men ages 20 and older is also created. When producing estimates for population subgroups (such as men 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 estimation of variance. 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: tab* command produces a cross-tabulation of HISPRACE2 and biokidsx and provides estimates appropriate to the complex sample design identified by the *svyset* command. The requested estimates and output are limited by specifying row, percent, and se after the *svy* command.

STATA Program

```
use "EX3.DTA"

svyset [pweight=WGT2017_2019], strata(SEST) psu(SECU)

generate biokidsx=0
replace biokidsx=1 if BIOKIDS>0

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

svy, subpop(agepop) row percent se: tab HISPRACE2 biokidsx
```

STATA Output



```
. svy, subpop(agepop) row percent se: tab HISPRACE2 biokidsx
 (running tabulate on estimation sample)

 Number of strata   =         18            Number of obs      =      5,206
 Number of PSUs     =         72            Population size    = 72,221,885
                                            Subpop. no. obs    =      4,174
                                            Subpop. size       = 62,331,053
                                            Design df          =         54

 ------------------------------------
 Race &     |
 Hispanic   |
 origin of  |
 responden  |
 t - 1997   |
 OMB        |
 standards  |       biokidsx
 (RECODE)   |       0        1   Total
 -----------+------------------------
  Hispanic  |   41.75    58.25     100
            |  (2.41)   (2.41)
            |
  Non-Hisp  |   50.05    49.95     100
            | (2.535)  (2.535)
            |
  Non-Hisp  |   45.98    54.02     100
            | (3.226)  (3.226)
            |
  Non-Hisp  |   58.29    41.71     100
            | (4.598)  (4.598)
            |
     Total  |   48.68    51.32     100
            | (2.116)  (2.116)
 ------------------------------------
   Key:  row percentage
         (linearized standard error of row percentage)

   Pearson:
     Uncorrected   chi2(3)         =   45.7307
     Design-based  F(2.89, 156.14) =    4.9479     P = 0.0030
```