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 2015-2017 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 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 SURVEYFREQ step produces a cross-tabulation of unweighted and weighted cell counts for the variables HISPRACE2 by biokidsx specified in the TABLE statement. The WEIGHT statement identifies the weight variable WGT2015_2017. 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. SAS documentation can provide more information about the NOMCAR option and options in the TABLE statement.

SAS Program

```
data EX3;
set NSFG.MALES;

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

**create a variable for your 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 WGT2015_2017;
run;
```

SAS Output

```
NSFG 2015-2017 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          4540
Sum of Weights              71617244


      Variance Estimation

Method            Taylor Series
Missing Values           NOMCAR

The SURVEYFREQ Procedure

                                            Table of HISPRACE2 by biokidsx
                                              Controlling for agepop=no

                                                                      Weighted    Std Err of        Row     Std Err of
HISPRACE2                              biokidsx        Frequency     Frequency      Wgt Freq    Percent    Row Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Hispanic                               none                  258       2346006        280348    99.4297         0.4418
                                       one or more             2         13457         10104     0.5703         0.4418

                                       Total                 260       2359463        278894    100.000
----------------------------------------------------------------------------------------------------------------------
Non-Hispanic White, Single Race        none                  357       5097877        481026    99.8524         0.1479
                                       one or more             1          7535          7535     0.1476         0.1479

                                       Total                 358       5105411        481134    100.000
----------------------------------------------------------------------------------------------------------------------
Non-Hispanic Black, Single Race        none                  179       1393079        187882    99.7931         0.2093
                                       one or more             1          2888          2888     0.2069         0.2093

                                       Total                 180       1395968        187806    100.000
----------------------------------------------------------------------------------------------------------------------
Non-Hispanic Other or Multiple Race    none                   87       1059345        162085    99.3924         0.6092
                                       one or more             1          6476          6476     0.6076         0.6092

                                       Total                  88       1065821        162336    100.000
----------------------------------------------------------------------------------------------------------------------
Total                                  none                  881       9896308        619753
                                       one or more             5         30355         14462

                                       Total                 886       9926663        618184
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

```
The SURVEYFREQ Procedure

                                            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
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Hispanic                               none                  306       5129945        645760    40.0914         2.5669
                                       one or more           420       7665688        930272    59.9086         2.5669

                                       Total                 726      12795633       1426721    100.000
----------------------------------------------------------------------------------------------------------------------
Non-Hispanic White, Single Race        none                 1031      17964564       1611032    50.1942         2.1420
                                       one or more           837      17825561       1444773    49.8058         2.1420

                                       Total                1868      35790125       2647829    100.000
----------------------------------------------------------------------------------------------------------------------
Non-Hispanic Black, Single Race        none                  324       3282482        430633    44.9035         3.2804
                                       one or more           367       4027603        541788    55.0965         3.2804

                                       Total                 691       7310084        844584    100.000
----------------------------------------------------------------------------------------------------------------------
Non-Hispanic Other or Multiple Race    none                  217       3328034        470216    57.4320         4.5355
                                       one or more           152       2466705        327769    42.5680         4.5355

                                       Total                 369       5794739        595667    100.000
----------------------------------------------------------------------------------------------------------------------
Total                                  none                 1878      29705025       2243207
                                       one or more          1776      31985556       1691504

                                       Total                3654      61690581       3365390
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

STATA 14

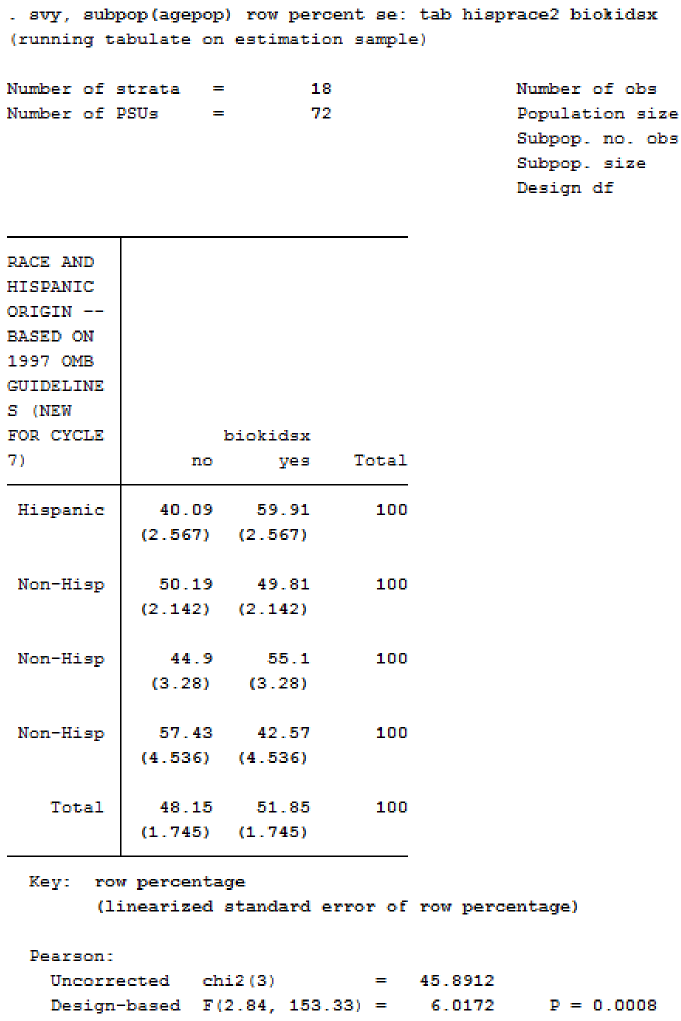
The use statement specifies the dataset to be used. The svyset command specifies the weight (WGT2015_2017), strata (SEST), and cluster (SECU) variables to be used in STATA in estimation. These settings are saved for the current session, but can be cleared by entering the clear command. The generate and 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 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: 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=WGT2015_2017], 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      =     4,540
Number of PSUs     =        72                Population size    = 71,617,244
                                              Subpop. no. obs    =     3,654
                                              Subpop. size       = 61,690,581
                                              Design df          =        54

-------------------------------------
RACE AND  |
HISPANIC  |
ORIGIN -- |
BASED ON  |
1997 OMB  |
GUIDELINE |
S (NEW    |
FOR CYCLE |         biokidsx
7)        |       no       yes    Total
----------+--------------------------
 Hispanic |    40.09     59.91      100
          |  (2.567)   (2.567)
          |
 Non-Hisp |    50.19     49.81      100
          |  (2.142)   (2.142)
          |
 Non-Hisp |     44.9      55.1      100
          |   (3.28)    (3.28)
          |
 Non-Hisp |    57.43     42.57      100
          |  (4.536)   (4.536)
          |
    Total |    48.15     51.85      100
          |  (1.745)   (1.745)
-------------------------------------
  Key:  row percentage
        (linearized standard error of row percentage)

  Pearson:
    Uncorrected   chi2(3)         =   45.8912
    Design-based  F(2.84, 153.33) =    6.0172     P = 0.0008
```