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

Percentage of Men 20-44 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 20-44 in the 2013-2015 NSFG data 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 upper case represent variables as named on the data files. Variables in lower case 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 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 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 WGT2013_2015. 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 percentages to the row

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;
stratum SEST;
cluster SECU;
table agepop*HISPRACE2*biokidsx / row;
weight WGT2013_2015;
run;
```

SAS Output

NSFG 2013-2015 Percentage of Males 20-44 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	4506
Sum of Weights	61157178

Table of HISPRACE2 by biokidsx
Controlling for agepop=no

HISPRACE2	biokidsx	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Hispanic	none	290	2266816	267291	22.8506	2.5341	98.9951	0.4555
	one or more	6	23010	10863	0.2319	0.1080	1.0049	0.4555
	Total	296	2289826	270142	23.0826	2.5565	100.000	
Non-Hispanic White, Single Race	none	417	5101625	397042	51.4269	2.6096	98.5446	0.8770
	one or more	5	75346	46059	0.7595	0.4606	1.4554	0.8770
	Total	422	5176972	402478	52.1864	2.6226	100.000	
Non-Hispanic Black, Single Race	none	168	1310371	169499	13.2092	1.6875	96.3258	2.8964
	one or more	3	49982	40611	0.5038	0.4096	3.6742	2.8964
	Total	171	1360353	175634	13.7130	1.7506	100.000	
Non-Hispanic Other or Multiple Race	none	108	1073285	146164	10.8192	1.3161	98.1954	1.2771
	one or more	2	19724	14322	0.1988	0.1429	1.8046	1.2771
	Total	110	1093009	148757	11.0181	1.3333	100.000	
Total	none	983	9752097	507036	98.3059	0.6261		
	one or more	16	168062	63154	1.6941	0.6261		
	Total	999	9920159	515066	100.000			

NSFG 2013-2015 Percentage of Males 20-44 Who Have Ever Fathered One or More Children by Hispanic Origin and Race

The SURVEYFREQ Procedure

Table of HISPRACE2 by biokidsx
Controlling for agepop=yes

HISPRACE2	biokidsx	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Hispanic	none	310	4399969	411069	8.5875	0.9058	40.6977	2.3651
	one or more	407	6411375	783010	12.5132	1.5062	59.3023	2.3651
	Total	717	10811344	1085432	21.1006	2.1931	100.000	
Non-Hispanic White, Single Race	none	994	14436523	1036972	28.1760	1.2653	49.0182	1.8734
	one or more	805	15014822	1430364	29.3046	1.8406	50.9818	1.8734
	Total	1799	29451345	2233532	57.4806	2.2959	100.000	
Non-Hispanic Black, Single Race	none	273	2607616	403333	5.0893	0.7438	43.2340	3.9642
	one or more	331	3423785	401297	6.6822	0.8347	56.7660	3.9642
	Total	604	6031401	653475	11.7716	1.2726	100.000	
Non-Hispanic Other or Multiple Race	none	227	2866169	382208	5.5939	0.7618	57.9852	3.5263
	one or more	160	2076761	337974	4.0532	0.6493	42.0148	3.5263
	Total	387	4942929	631744	9.6472	1.2366	100.000	
Total	none	1804	24310277	1236944	47.4467	1.4783		
	one or more	1703	26926742	1611096	52.5533	1.4783		
	Total	3507	51237019	2421335	100.000			

STATA 14

The use statement specifies the dataset to be used. The svyset command specifies the weight (WGT2013_2015), 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 HISPRACE 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=WGT2013_2015], 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      =       4,506
Number of PSUs     =        72                  Population size    =  61,157,178
                                                Subpop. no. obs    =       3,507
                                                Subpop. size       =  51,237,019
                                                Design df          =          54
```

Race & Hispanic origin of respondent - 1997 OMB standards (RECODE)	biokidsx no	biokidsx yes	Total
Hispanic	40.7	59.3	100
	(2.365)	(2.365)	
Non-Hisp	49.02	50.98	100
	(1.873)	(1.873)	
Non-Hisp	43.23	56.77	100
	(3.964)	(3.964)	
Non-Hisp	57.99	42.01	100
	(3.526)	(3.526)	
Total	47.45	52.55	100
	(1.478)	(1.478)	

```
  Key:  row percentage
        (linearized standard error of row percentage)

  Pearson:
    Uncorrected   chi2(3)         =   43.0707
    Design-based  F(2.91, 157.07) =    5.6935     P = 0.0011
```