Example 1: Variance Estimates for Percentages using SAS (9.4) and STATA (18)

Percentage of Women Ages 15-49 Currently Using the Oral Contraceptive Pill, by Age

Following are SAS and STATA programs and output for an analysis of the percentage of women in the 2022-2023 NSFG female respondent file who were using the oral contraceptive pill during the month of interview. A cross-tabulation of use of the pill by age (15-19, 20-24, 25-29, 30-34, and 40-49) is generated.

The estimates and standard errors calculated 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 their computing environment. Formatting and library options have been deleted 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, age categories (agerx), and current use of contraceptive pill (cpill). The PROC SURVEYFREQ produces a cross-tabulation of unweighted and weighted cell counts for the variables specified in the TABLE statement (agerx and cpill). The WEIGHT statement identifies the weight variable WGT2022_2023. PROC SURVEYFREQ calculates standard errors appropriate to the complex sample design identified 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 EX1;
set NSFG.FEMALES (keep=CASEID AGER CONSTAT1 VEST VECL WGT2022_2023);

if 15 le AGER le 19 then agerx=1;
else if 20 le AGER le 24 then agerx=2;
else if 25 le AGER le 29 then agerx=3;
else if 30 le AGER le 34 then agerx=4;
else if 35 le AGER le 39 then agerx=5;
else if AGER ge 40 then agerx=6;

**Value of 6 on CONSTAT1 is oral contraceptive pill;
if CONSTAT1=6 then cpill=1;
else cpill=2;
run;

proc surveyfreq nomcar;
stratum VEST;
cluster VECL;
weight WGT2022_2023;
table agerx*cpill /ROW NOCELLPERCENT nosparse;
run;
```

SAS Output

NSFG 2022-2023 Percentage of Women Using the Pill by Age

The SURVEYFREQ Procedure

Data Summary

Number of Strata	20
Number of Clusters	80
Number of Observations	5586
Sum of Weights	74936917.9

Variance Estimation

Method	Taylor Series
Missing Values	NOMCAR

The SURVEYFREQ Procedure

Table of agerx by cpill

agerx	cpill	Frequency	Weighted Frequency	Std Err of Wgt Freq	Row Percent	Std Err of Row Percent
15-19	yes	96	1493243	207476	14.2348	1.6792
	no	634	8996818	469101	85.7652	1.6792
	Total	730	10490061	552116	100.0000	
20-24	yes	111	2071629	232419	18.9586	2.0226
	no	495	8855484	546691	81.0414	2.0226
	Total	606	10927114	579466	100.0000	
25-29	yes	116	1562548	217882	14.6057	1.8889
	no	687	9135628	621945	85.3943	1.8889
	Total	803	10698176	671258	100.0000	
30-34	yes	113	1158344	170778	10.1973	1.3836
	no	915	10200933	641461	89.8027	1.3836
	Total	1028	11359278	688812	100.0000	
35-39	yes	79	857487	124595	7.8114	1.1050
	no	849	10119939	554957	92.1886	1.1050
	Total	928	10977426	573133	100.0000	
40-49	yes	118	1405912	179174	6.8632	0.7961
	no	1373	19078951	1004109	93.1368	0.7961
	Total	1491	20484864	1065014	100.0000	
Total	yes	633	8549165	568019		
	no	4953	66387753	2651588		

The SURVEYFREQ Procedure

Table of agerx by cpill

agerx	cpill	Frequency	Weighted Frequency	Std Err of Wgt Freq	Row Percent	Std Err of Row Percent
Total	Total	5586	74936918	2910451		

STATA 18

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (WGT2022_2023), strata (VEST), and cluster (VECL) variables to be used by STATA in estimation. These settings are saved for the current session but can be cleared by entering the *clear* command or running *svyset* again with different settings. The *generate* and *replace* statements create the recoded variables agerx and cpill. The *svytab* command produces a cross-tabulation of agerx and pill and provides estimates appropriate to the complex sample design identified by the *svyset* command. The requested estimates and output are limited by specifying *row* and *se* after the *svytab* command.

STATA Program

```
use "EX1.DTA"

svyset [pweight=WGT2022_2023], strata(VEST) psu(VECL)

generate agerx=1 if AGER <=19
replace agerx=2 if AGER >=20 & AGER <=24
replace agerx=3 if AGER >=25 & AGER <=29
replace agerx=4 if AGER >=30 & AGER <=34
replace agerx=5 if AGER >=35 & AGER <=39
replace agerx=6 if AGER >=40

generate cpill=2
replace cpill=1 if CONSTAT1==6

svy: tab agerx cpill, row se percent
```

STATA Output

```
. svy: tab agerx cpill, row se percent
(running tabulate on estimation sample)

Number of strata = 20                        Number of obs   =      5,586
Number of PSUs   = 80                        Population size = 74,936,918
                                             Design df       =         60
```

agerx	cpill: yes	cpill: no	Total
15-19	14.23	85.77	100
	(1.679)	(1.679)	
20-24	18.96	81.04	100
	(2.023)	(2.023)	
25-29	14.61	85.39	100
	(1.889)	(1.889)	
30-34	10.2	89.8	100
	(1.384)	(1.384)	
35-39	7.811	92.19	100
	(1.105)	(1.105)	
40-49	6.863	93.14	100
	(.7961)	(.7961)	
Total	11.41	88.59	100
	(.6425)	(.6425)	

```
Key: Row percentage
     (Linearized standard error of row percentage)

  Pearson:
    Uncorrected   chi2(5)         =  103.1057
    Design-based  F(4.32, 258.92) =   12.0408     P = 0.0000
```

Example 2: Variance Estimates for Means using SAS (9.4) and STATA (18)

Mean Number of Children Ever Born, by Urban/Rural Residence for Women 15-49 Years of Age

Following are SAS and STATA programs and output for an analysis of the mean number of children born to women 15-49 years of age in the 2022-2023 NSFG female respondent file, by urban/rural residence.

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

In these programs, library and file names are generic; the user must apply names specific to their computing environment. Formatting and library options are not presented since preferences will vary across user organizations.

SAS 9.4

The DATA step creates a dataset for females that contains the variables to be used in the analysis. The PROC SURVEYMEANS step produces a table of weighted means for the variable specified in the VAR statement (PARITY) by urban/rural residence (urban) by using the DOMAIN statement. The WEIGHT statement identifies the weight variable (WGT2022_2023) 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. Data users should consult official SAS documentation for more information about the NOMCAR option.

SAS Program

```
data NSFG.EX2;
set NSFG.FEMALES (keep=CASEID VEST VECL METRO PARITY WGT2022_2023);

if METRO in (1,2) then urban=1;
else if METRO eq 3 then urban=2;

run;

proc surveymeans nomcar;
stratum VEST;
cluster VECL;
domain urban;
var PARITY;
weight WGT2022_2023;
run;
```

SAS Output

NSFG 2022-2023 Mean Numbers of Children Ever Born (PARITY) by urban/rural residence

The SURVEYMEANS Procedure

Data Summary

Number of Strata	20
Number of Clusters	80
Number of Observations	5586
Sum of Weights	74936917.9

Variance Estimation

Method	Taylor Series
Missing Values	NOMCAR

Statistics

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
PARITY	Number of live births	5586	1.108647	0.031077	1.04648418	1.17080896

Statistics for urban Domains

urban	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
urban	PARITY	Number of live births	4796	1.056112	0.033180	0.98974177	1.12248216
rural	PARITY	Number of live births	790	1.384746	0.098267	1.18818322	1.58130789

STATA 14.0

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (WGT2022_2023), strata (VEST), and cluster (VECL) variables to be used by STATA in estimation. These settings are saved for the current session but can be cleared by entering the *clear* command.

The *svy: mean* command produces estimated weighted means for each of the levels of the by variable metro to show means separately by urban/rural 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.

STATA Program

```
use "EX2.DTA"

generate urban=1
replace urban=2 if METRO==3

svyset [pweight=WGT2022_2023], strata(VEST) psu(VECL)

svy: mean PARITY, over(metro)
```

STATA Output

```
. svy: mean PARITY, over(urban)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 20              Number of obs   =      5,586
Number of PSUs   = 80              Population size = 74,936,918
                                   Design df       =         60
```

	Mean	Linearized std. err.	[95% conf.	interval]
c.PARITY@urban				
urban	1.056112	.0331802	.9897418	1.122482
rural	1.384746	.0982665	1.188183	1.581308

```
.
```

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 2022-2023 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 variable indicating whether the respondent fathered one or more biological children (biokidsx) based on the variable EVBIOKID. For this example, respondents who said 'don't know' or refused to answer EVBIOKID are coded as missing (sysmis) on biokidsx, but analysts may have different approaches. 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 WGT2022_2023. 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 VEST VECL WGT2022_2023);

biokidsx=0;
if EVBIOKID eq 1 then biokidsx=1;
else if EVBIOKID in (8 9) then biokidsx=.;

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

proc surveyfreq nomcar;
stratum VEST;
cluster VECL;
table agepop*HISPRACE2*biokidsx / ROW NOCELLPERCENT nosparse;
weight WGT2022_2023;
run;
```

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

NSFG 2022-2023 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	20
Number of Clusters	80
Number of Observations	4371
Sum of Weights	75700206.4

Variance Estimation

Method	Taylor Series
Missing Values	NOMCAR

Table of HISPRACE2 by biokidsx

Controlling for agepop=yes

HISPRACE2	biokidsx	Frequency	Weighted Frequency	Std Err of Wgt Freq	Row Percent	Std Err of Row Percent
Hispanic	none	406	7190052	823443	51.4311	2.5095
	one or more	266	6789912	909279	48.5689	2.5095
	Total	672	13979964	1589950	100.0000	
Non-Hispanic White, Single Race	none	1229	19299122	1203242	55.0220	1.7217
	one or more	797	15776165	1079723	44.9780	1.7217
	Total	2026	35075287	1936390	100.0000	
Non-Hispanic Black, Single Race	none	245	3727842	391396	49.6898	3.0161
	one or more	188	3774383	400928	50.3102	3.0161
	Total	433	7502225	650372	100.0000	
Non-Hispanic Other or Multiple Race	none	316	4435667	674037	56.7676	3.3729
	one or more	197	3378066	561211	43.2324	3.3729
	Total	513	7813733	1116169	100.0000	
Total	none	2196	34652683	1755313		
	one or more	1448	29718526	1521689		
	Total	3644	64371210	2865200		

STATA 14

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (WGT2022_2023), strata (VEST), and cluster (VECL) 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 based on the variable EVBIOKID. For this example, respondents who said don't know or refused to answer EVBIOKID are coded as missing (sysmis) on biokidsx, but analysts may have different approaches. 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=WGT2022_2023], strata(VEST) psu(VECL)

generate biokidsx=0
replace biokidsx=1 if EVBIOKID==1
replace biokidsx=. if EVBIOKID==8
replace biokidsx=. if EVBIOKID==9

* 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 = 20                      Number of obs   =      4,343
Number of PSUs   = 80                      Population size = 75,281,470
                                           Subpop. no. obs =      3,644
                                           Subpop. size    = 64,371,210
                                           Design df       =         60

------------------------------------
Race and  |
Hispanic  |
origin -  |
based on  |
1997 OMB  |
guideline |          biokidsx
s         |     none       1+    Total
----------+-------------------------
        1 |    51.43    48.57      100
          |   (2.51)   (2.51)
          |
        2 |    55.02    44.98      100
          |  (1.722)  (1.722)
          |
        3 |    49.69    50.31      100
          |  (3.016)  (3.016)
          |
        4 |    56.77    43.23      100
          |  (3.373)  (3.373)
          |
    Total |    53.83    46.17      100
          |  (1.232)  (1.232)
------------------------------------
Key: Row percentage
     (Linearized standard error of row percentage)

  Pearson:
    Uncorrected   chi2(3)         =    8.8582
    Design-based  F(2.84, 170.34) =    1.2984     P = 0.2770

.
```