Example 1: Variance Estimates for Percentages using SAS (9.4) and STATA (14)

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 2015-2017 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 (in six categories: 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 his or her 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 (i.e. agerx by cpill) 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 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. SAS documentation can provide more information about the NOMCAR option and options in the TABLE statement.

SAS Program

```
data EX1;
set NSFG.FEMALES;

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 SEST;
cluster SECU;
weight WGT2015_2017;
table agerx*cpill /ROW NOCELLPERCENT nosparse;
run;
```

SAS Output

NSFG 2015-2017 Percentage of Women Using the Pill by Age

The SURVEYFREQ 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

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	111	1573631	272171	16.6446	2.4417
	no	813	7880685	711205	83.3554	2.4417
	Total	924	9454316	817177	100.000	
20-24	yes	153	2289761	327074	22.5002	2.4761
	no	591	7886847	718730	77.4998	2.4761
	Total	744	10176608	881972	100.000	
25-29	yes	144	1861427	256837	16.6835	1.7382
	no	790	9295897	684021	83.3165	1.7382
	Total	934	11157324	828488	100.000	
30-34	yes	119	1509510	256188	14.0482	2.0045
	no	760	9235708	732586	85.9518	2.0045
	Total	879	10745218	841067	100.000	
35-39	yes	68	807776	160633	7.8170	1.5849
	no	684	9525811	762372	92.1830	1.5849
	Total	752	10333587	767041	100.000	
40-49	yes	77	1047084	194820	5.1451	0.9642
	no	1244	19303949	1177412	94.8549	0.9642
	Total	1321	20351033	1187512	100.000	
Total	yes	672	9089188	737224		

Table of agerx by cpill

agerx	cpill	Frequency	Weighted Frequency	Std Err of Wgt Freq	Row Percent	Std Err of Row Percent
Total	no	4882	63128898	3298847		
	Total	5554	72218086	3791652		

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 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=WGT2015_2017], strata(SEST) psu(SECU)

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

agerx	cpill: yes	cpill: no	Total
15-19	16.64	83.36	100
	(2.442)	(2.442)	
20-24	22.5	77.5	100
	(2.476)	(2.476)	
25-29	16.68	83.32	100
	(1.738)	(1.738)	
30-34	14.05	85.95	100
	(2.004)	(2.004)	
35-39	7.817	92.18	100
	(1.585)	(1.585)	
40-49	5.145	94.85	100
	(.9642)	(.9642)	
Total	12.59	87.41	100
	(.7102)	(.7102)	

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

Pearson:
  Uncorrected   chi2(5)         =  190.7037
  Design-based  F(4.71, 254.25) =   13.7941     P = 0.0000
```