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 2017-2019 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 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 specified in the TABLE statement (agerx and cpill). The WEIGHT statement identifies the weight variable WGT2017_2019. 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 SEST SECU WGT2017_2019);

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 WGT2017_2019;
table agerx*cpill /ROW NOCELLPERCENT nosparse;
run;
```

SAS Output

NSFG 2017-2019 Percentage of Women Currently Using the Pill by Age

The SURVEYFREQ Procedure

Data Summary

Number of Strata	18
Number of Clusters	72
Number of Observations	6141
Sum of Weights	72671926

Variance Estimation

Method	Taylor Series
Missing Values	NOMCAR

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	160	1841692	200284	19.5112	1.8670
	no	810	7597485	556481	80.4888	1.8670
	Total	970	9439177	626554	100.0000	
20-24	yes	173	2366685	308996	23.7833	2.1713
	no	639	7584340	604905	76.2167	2.1713
	Total	812	9951025	789942	100.0000	
25-29	yes	163	2252838	330322	19.6916	2.2773
	no	853	9187786	607161	80.3084	2.2773
	Total	1016	11440624	764140	100.0000	
30-34	yes	141	1661147	218936	15.2800	1.7551
	no	903	9210267	562952	84.7200	1.7551
	Total	1044	10871414	637599	100.0000	
35-39	yes	66	695757	107752	6.4965	0.9895
	no	789	10013957	677843	93.5035	0.9895
	Total	855	10709714	696600	100.0000	
40-49	yes	84	1321257	220073	6.5215	1.0029
	no	1360	18938715	1178194	93.4785	1.0029
	Total	1444	20259972	1245158	100.0000	
Total	yes	787	10139377	778711		
	no	5354	62532549	3087393		

Table of agerx by cpill

agerx	cpill	Frequency	Weighted Frequency	Std Err of Wgt Freq	Row Percent	Std Err of Row Percent
Total	Total	6141	72671926	3521465		

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 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=WGT2017_2019], 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      =        6,141
Number of PSUs     =        72                  Population size    =   72,671,926
                                                Design df          =           54
```

agerx	cpill: yes	cpill: no	Total
15-19	19.51	80.49	100
	(1.867)	(1.867)	
20-24	23.78	76.22	100
	(2.171)	(2.171)	
25-29	19.69	80.31	100
	(2.277)	(2.277)	
30-34	15.28	84.72	100
	(1.755)	(1.755)	
35-39	6.497	93.5	100
	(.9895)	(.9895)	
40-49	6.522	93.48	100
	(1.003)	(1.003)	
Total	13.95	86.05	100
	(.829)	(.829)	

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

Pearson:
  Uncorrected   chi2(5)         =  236.7424
  Design-based  F(4.39, 236.89) =   23.0013     P = 0.0000
```