Example 2: Variance estimates for Means using SAS (9.4) and STATA (14)

Mean Number of Children Ever Born, by Hispanic Origin and Race for Women 20-44 Years of Age

Following are SAS and STATA programs and output for an analysis of the mean number of children born to women 20-44 years of age in the 2013-2015 NSFG data file, 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 for females which contains the variables to be used in the analysis and the subpopulation indicator variable (agepop) that is used to identify women ages 20 and older. When producing estimates for population subgroups (such as women 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 SURVEYMEANS step produces a table of weighted means for the variable specified in the VAR statement (PARITY) by Hispanic Origin and Race (HISPRACE2) by using the DOMAIN statement. The WEIGHT statement identifies the weight variable (WGT2013_2015) 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.

SAS Program

```
data NSFG.EX2;
set NSFG.FEMALES;

**Create a variable for your subpopulation of ages 20 and older;
agepop=0;
if ager ge 20 then agepop=1;
run;

proc surveymeans;
stratum sest;
cluster secu;
domain agepop*hisprace2;
var parity;
weight WGT2013_2015;
run;
```

SAS Output

Mean Numbers of Children Ever Born (PARITY) by Hispanic Origin and Race for Women Ages 20-44

The SURVEYMEANS Procedure

Data Summary

Number of Strata	18
Number of Clusters	72
Number of Observations	5699
Sum of Weights	61491766

Statistics

Variable	Label	N	Mean	Std Error of Mean
PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)	5699	1.171290	0.033611

Statistics

Variable	95% CL for Mean	
PARITY	1.10390413	1.23867509

Mean Numbers of Children Ever Born (PARITY) by Hispanic Origin and Race for Women Ages 20-44

The SURVEYMEANS Procedure

Domain Statistics in agepop*HISPRACE2

agepop	Race & Hispanic origin of respondent - 1997 OMB standards (RECODE)	Variable	Label
yes	Hispanic	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
	Non-Hispanic White, Single Race	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
	Non-Hispanic Black, Single Race	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
	Non-Hispanic Other or Multiple Race	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
no	Hispanic	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
	Non-Hispanic White, Single Race	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
	Non-Hispanic Black, Single Race	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)
	Non-Hispanic Other or Multiple Race	PARITY	CAPI-based total number of live births (accounting for mult birth) (RECODE)

Domain Statistics in agepop*HISPRACE2

agepop	Race & Hispanic origin of respondent - 1997 OMB standards (RECODE)	Variable	N	Mean	Std Error of Mean	95% CL for Mean	
yes	Hispanic	PARITY	1071	1.751282	0.056829	1.6373474	1.86521664
	Non-Hispanic White, Single Race	PARITY	2181	1.263799	0.053881	1.1557748	1.37182416
	Non-Hispanic Black, Single Race	PARITY	936	1.508209	0.094774	1.3181975	1.69822024
	Non-Hispanic Other or Multiple Race	PARITY	501	1.120707	0.083557	0.9531855	1.28822928
no	Hispanic	PARITY	323	0.052835	0.013592	0.0255853	0.08008441
	Non-Hispanic White, Single Race	PARITY	408	0.023193	0.009572	0.0040022	0.04238347
	Non-Hispanic Black, Single Race	PARITY	178	0.048379	0.030174	-0.0121169	0.10887528
	Non-Hispanic Other or Multiple Race	PARITY	101	0.023415	0.012652	-0.0019504	0.04877993

STATA 14.0

STATA Program

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 by 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 indicating the subpopulation of women ages 20 and older. When producing estimates for population subgroups (such as women 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: mean* command produces estimated weighted means for each of the levels of the by variable HISPRACE2 to show means separately by Hispanic origin and race 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.

```
use "EX2.DTA"

svyset [pweight=WGT2013_2015], strata(SEST) psu(SECU)

* create a variable for your subpopulation of ages 20 and older
generate agepop=0
replace agepop=1 if AGER>=20

svy: mean parity, over(agepop hisprace2)
```

STATA Output

```
. svy: mean parity, over(agepop hisprace2)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      18        Number of obs    =       5,699
Number of PSUs   =      72        Population size  =  61,491,766
                                  Design df        =          54

         Over: agepop hisprace2
    _subpop_1: yes Hispanic
    _subpop_2: yes Non-Hispanic White, Single R
    _subpop_3: yes Non-Hispanic Black, Single R
    _subpop_4: yes Non-Hispanic Other or Multip
    _subpop_5: no Hispanic
    _subpop_6: no Non-Hispanic White, Single Ra
    _subpop_7: no Non-Hispanic Black, Single Ra
    _subpop_8: no Non-Hispanic Other or Multipl
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
parity				
_subpop_1	1.751282	.0568287	1.637347	1.865217
_subpop_2	1.263799	.0538809	1.155775	1.371824
_subpop_3	1.508209	.0947745	1.318198	1.69822
_subpop_4	1.120707	.0835571	.9531855	1.288229
_subpop_5	.0528349	.0135916	.0255853	.0800844
_subpop_6	.0231928	.009572	.0040022	.0423835
_subpop_7	.0483792	.0301744	-.0121169	.1088753
_subpop_8	.0234148	.0126517	-.0019504	.0487799