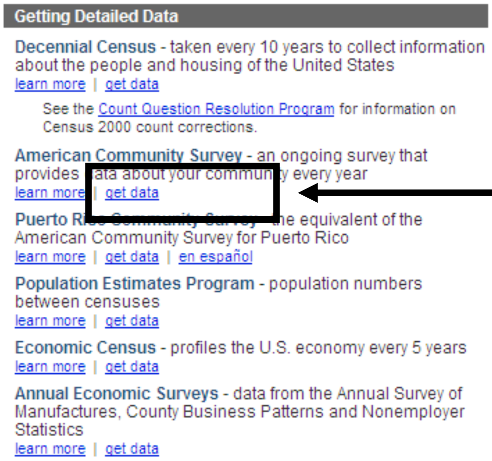
Center for Family and Demographic Research

December 2009

Working with the American Community Survey PUMS Data: Understanding and Using Replicate Weights

Step 1: Accessing the Data

The easiest way to attain estimates from the American Community Survey (ACS) is through American Fact Finder http://factfinder.census.gov/home/saff/main.html?_lang=en. However, sometimes researchers need information that is not published on the American Fact Finder tables. In this case, researchers need to use the Public Use Microdata Sample files (PUMS) for their analysis. First, the correct dataset must be obtained. The easiest way to download the data is through American Fact Finder at the web address above.



Getting Detailed Data

Decennial Census - taken every 10 years to collect information about the people and housing of the United States
learn more | get data

See the Count Question Resolution Program for information on Census 2000 count corrections.

American Community Survey - an ongoing survey that provides data about your community every year
learn more | get data

Puerto Rico Community Survey - the equivalent of the American Community Survey for Puerto Rico
learn more | get data | en español

Population Estimates Program - population numbers between censuses
learn more | get data

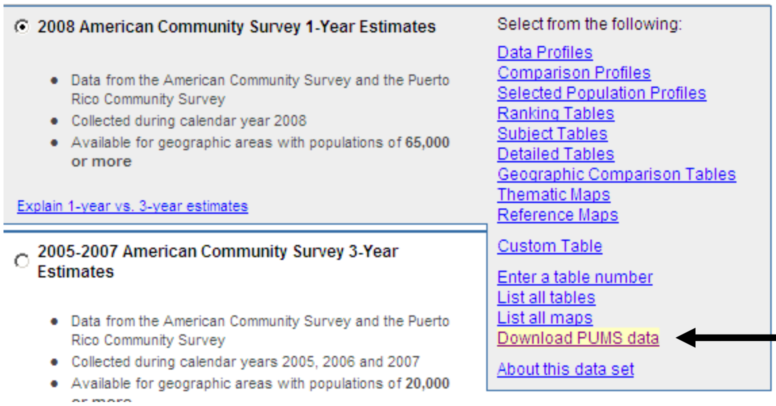
Economic Census - profiles the U.S. economy every 5 years
learn more | get data

Annual Economic Surveys - data from the Annual Survey of Manufactures, County Business Patterns and Nonemployer Statistics
learn more | get data

Click here to access the ACS datasets.

Next, pick the ACS dataset that is right for the research question. These datasets are very large, so it is very important to think before downloading. Both single year estimates and three year estimates are available. The single year estimates are used for populations of 65,000 or more and three year estimates are used for populations of 20,000 or more. The U.S. Census Bureau releases informative handbooks on the ACS called the Compass Series which can help determine if the research question should be answered with one year estimates or three year estimates (and soon to be released five year estimates). These handbooks are useful including one called, “What Researchers Need to Know” found here: http://www.census.gov/acs/www/Downloads/ACSResearch.pdf

We suggest this document is reviewed before using the ACS PUMS datasets.



● 2008 American Community Survey 1-Year Estimates

- Data from the American Community Survey and the Puerto Rico Community Survey
- Collected during calendar year 2008
- Available for geographic areas with populations of **65,000 or more**

Explain 1-year vs. 3-year estimates

○ 2005-2007 American Community Survey 3-Year Estimates

- Data from the American Community Survey and the Puerto Rico Community Survey
- Collected during calendar years 2005, 2006 and 2007
- Available for geographic areas with populations of **20,000**

Select from the following:

Data Profiles
Comparison Profiles
Selected Population Profiles
Ranking Tables
Subject Tables
Detailed Tables
Geographic Comparison Tables
Thematic Maps
Reference Maps

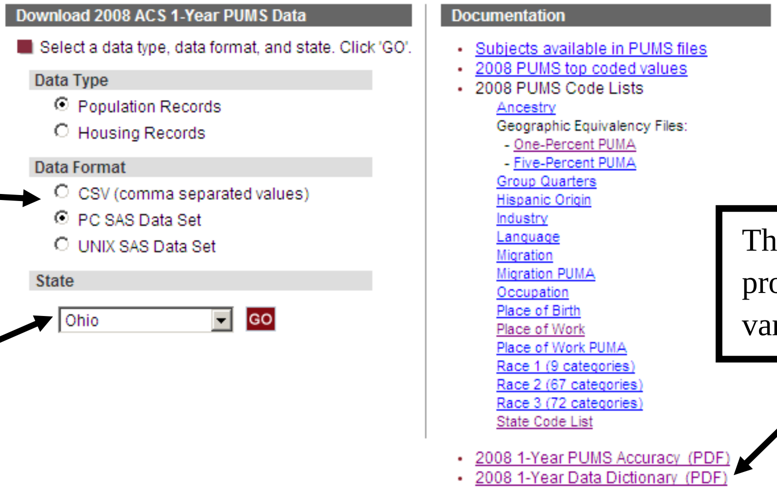
Custom Table

Enter a table number
List all tables
List all maps
Download PUMS data

About this data set

Click here to down load the PUMS data file.

What Type of data do you want: population or housing?



Download 2008 ACS 1-Year PUMS Data

Select a data type, data format, and state. Click 'GO'.

Data Type

● Population Records
○ Housing Records

Data Format

○ CSV (comma separated values)
● PC SAS Data Set
○ UNIX SAS Data Set

State

Ohio GO

Documentation

- Subjects available in PUMS files
- 2008 PUMS top coded values
- 2008 PUMS Code Lists
  Ancestry
  Geographic Equivalency Files:
  - One-Percent PUMA
  - Five-Percent PUMA
  Group Quarters
  Hispanic Origin
  Industry
  Language
  Migration
  Migration PUMA
  Occupation
  Place of Birth
  Place of Work
  Place of Work PUMA
  Race 1 (9 categories)
  Race 2 (67 categories)
  Race 3 (72 categories)
  State Code List
- 2008 1-Year PUMS Accuracy (PDF)
- 2008 1-Year Data Dictionary (PDF)

The ACS is available in CSV, SAS, or UNIX files.

What geographical location do you want: national or state level?

The data dictionary provides a list of the variables.

It is important to include weights in your data file. Point estimates require the person weight (pwgtp) or household weight (wgtp). To estimate standard errors, both the person (or household) weight and the corresponding replicate weights, pwgtp1-pwgtp80 for persons or wgtp1-wgtp80 for households, are required.

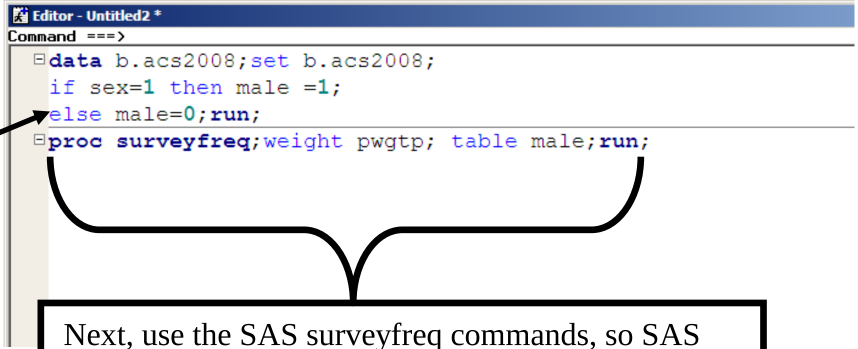
Step 2: Population Estimates

This handout will show you how to analyze ACS PUMS data using both SAS and STATA. As shown above, the data can be downloaded from the U.S. Census Bureau's webpage as a SAS file. This handout will first show you how to get estimates using SAS then followed by an example using STATA. This document will then show you how to get the correct standard errors, so you can perform hypothesis tests. For analysis, two types of weights are needed: person weight (or household) and the replicate weights. The person weight is required for the point estimates and both person weight and the replicate weights are necessary to calculate correct standard errors.

Let's say we want to know how many males are in Ohio using the 2008 ACS PUMS file. The Ohio Individual file SAS dataset should be downloaded and opened in SAS.

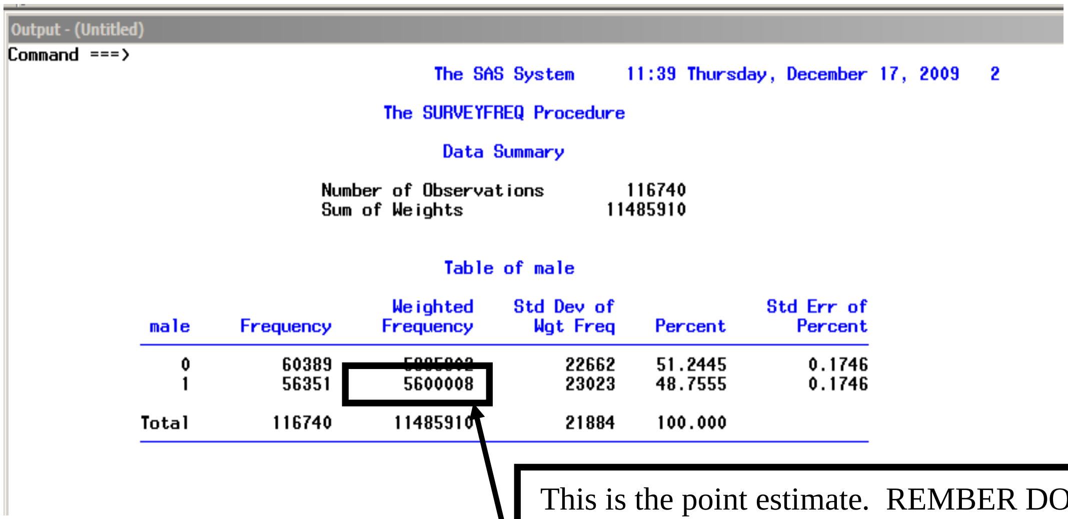
Using SAS:

First, you create a new variable where "1" is the sample you want to analyze and "0" equals everyone else in the sample.



```
Command ===>
data b.acs2008;set b.acs2008;
if sex=1 then male =1;
else male=0;run;
proc surveyfreq;weight pwgtp; table male;run;
```

Next, use the SAS surveyfreq commands, so SAS knows to use the weight in the estimate. Then in SAS use the person weight "pwgtp." Finally, run the frequency on the newly created variable in this case "male."



```
Command ===>
                                  The SAS System      11:39 Thursday, December 17, 2009   2

                              The SURVEYFREQ Procedure

                                   Data Summary

                      Number of Observations          116740
                      Sum of Weights                11485910


                                   Table of male

                                 Weighted    Std Dev of                  Std Err of
     male        Frequency      Frequency      Wgt Freq       Percent       Percent
     ------------------------------------------------------------------------------
        0            60389        5885902         22662       51.2445        0.1746
        1            56351        5600008         23023       48.7555        0.1746

     Total          116740       11485910         21884       100.000
     ------------------------------------------------------------------------------
```

This is the point estimate. REMBER DO NOT USE THE STD IN THIS OUTPUT. This handout will show you how to get the correct standard error.

Using STATA:

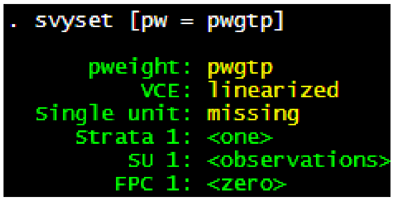
There is not an option to download the ACS PUMS data as a STATA data format, so use the program StatTransfer to change the file from a SAS sas.7bdat file into a STATA .dta file. Since these files are large, you may have to increase the memory size in STATA. For example:

```
set mem 200m
```

This will increase the memory your computer allows STATA to use. Open the dataset in STATA. In addition, selection of variables used in analyses will conserve space. Make sure the eighty-one weight variables are included, for example the person weights: pwgtp pwgtp1-pwgtp80. DO NOT DELETE ANY INDIVIDUAL CASES.

Now set up STATA, so that it is able to use the person weights needed to perform the estimates.

```
svyset [pw = pwgtp]
```



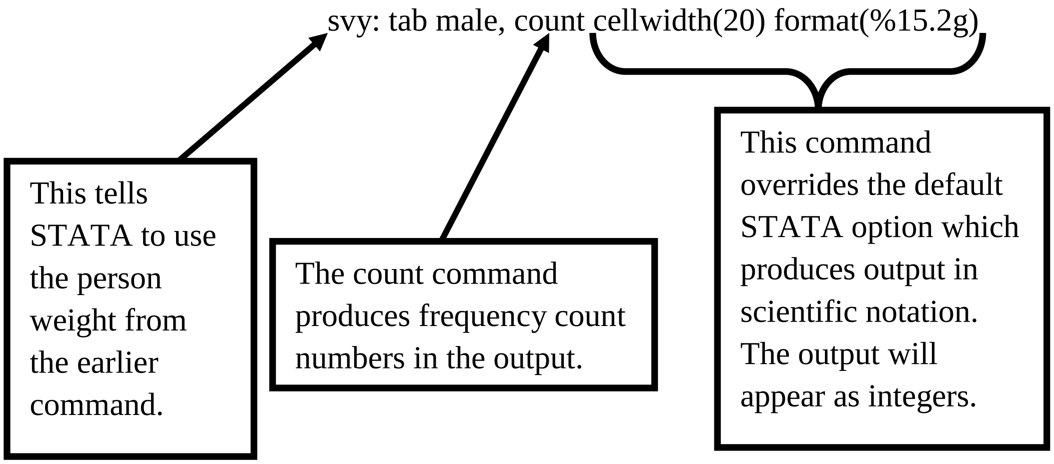
```
. svyset [pw = pwgtp]

      pweight: pwgtp
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

Your results file should look like this and conveys to STATA that weights will be used in analyses.

Create a variable where “1”=Males and “0”=everyone else.

```
gen male=0
replace male=1 if sex==1
```

To use the person weight, the svy command must be included in the syntax as follows. For more help on the svy command, type in the command window “help svy.”

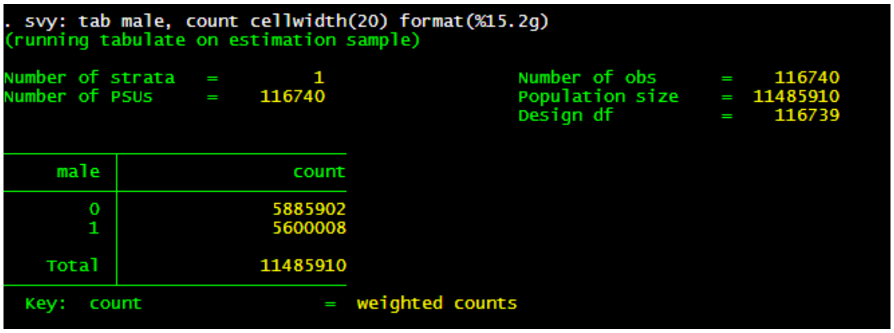


```
svy: tab male, count cellwidth(20) format(%15.2g)
```

This tells STATA to use the person weight from the earlier command.

The count command produces frequency count numbers in the output.

This command overrides the default STATA option which produces output in scientific notation. The output will appear as integers.

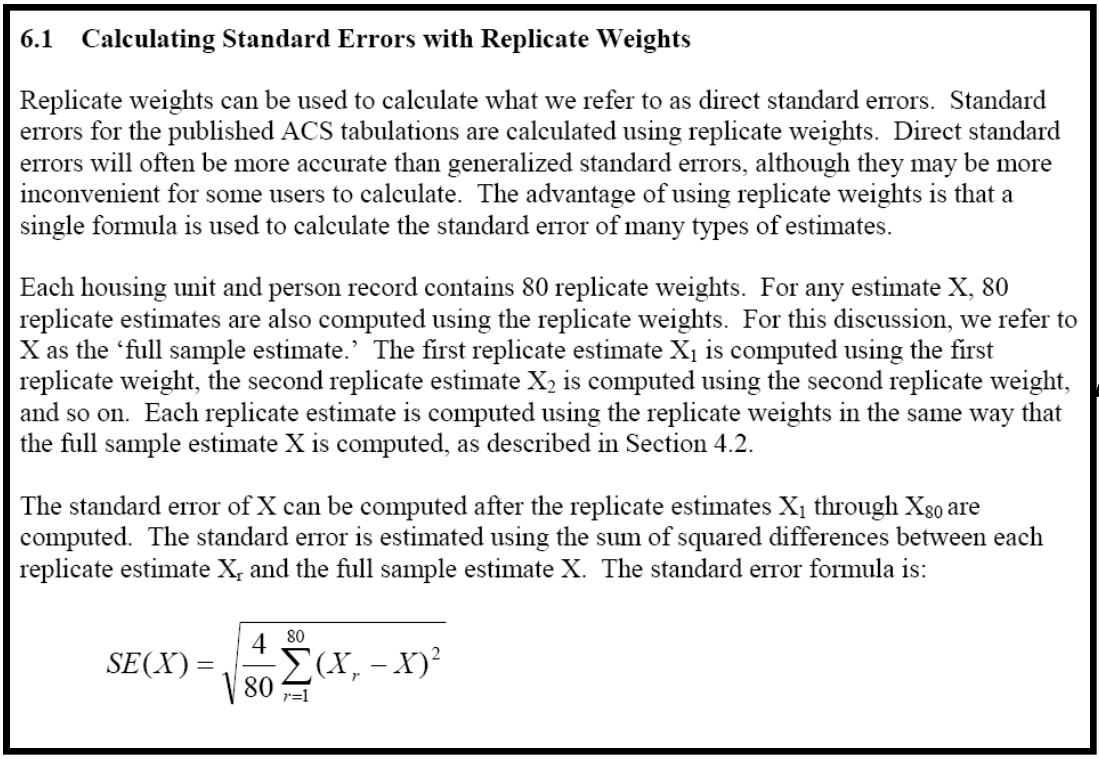


```
. svy: tab male, count cellwidth(20) format(%15.2g)
(running tabulate on estimation sample)

Number of strata   =         1                  Number of obs      =    116740
Number of PSUs     =    116740                  Population size    =  11485910
                                                Design df          =    116739

----------------------------------
     male |                 count
----------+-----------------------
        0 |               5885902
        1 |               5600008
          |
    Total |              11485910
----------------------------------
  Key:  count              =  weighted counts
```

The same results are obtained in STATA as in SAS.

Step 3: Calculating the Standard Errors.

Standard errors allow the calculation of confidence intervals, coefficients of variation, and significance tests. For more information and a detailed explanation of these statistics, please see the U.S. Census Bureau's Compass Series. The U.S. Census Bureau also provides an equation that produces the correct standard errors. The Accuracy Report can be found here: http://www.census.gov/acs/www/Downloads/2008/AccuracyPUMS.pdf.



6.1 Calculating Standard Errors with Replicate Weights

Replicate weights can be used to calculate what we refer to as direct standard errors. Standard errors for the published ACS tabulations are calculated using replicate weights. Direct standard errors will often be more accurate than generalized standard errors, although they may be more inconvenient for some users to calculate. The advantage of using replicate weights is that a single formula is used to calculate the standard error of many types of estimates.

Each housing unit and person record contains 80 replicate weights. For any estimate X, 80 replicate estimates are also computed using the replicate weights. For this discussion, we refer to X as the 'full sample estimate.' The first replicate estimate X_1 is computed using the first replicate weight, the second replicate estimate X_2 is computed using the second replicate weight, and so on. Each replicate estimate is computed using the replicate weights in the same way that the full sample estimate X is computed, as described in Section 4.2.

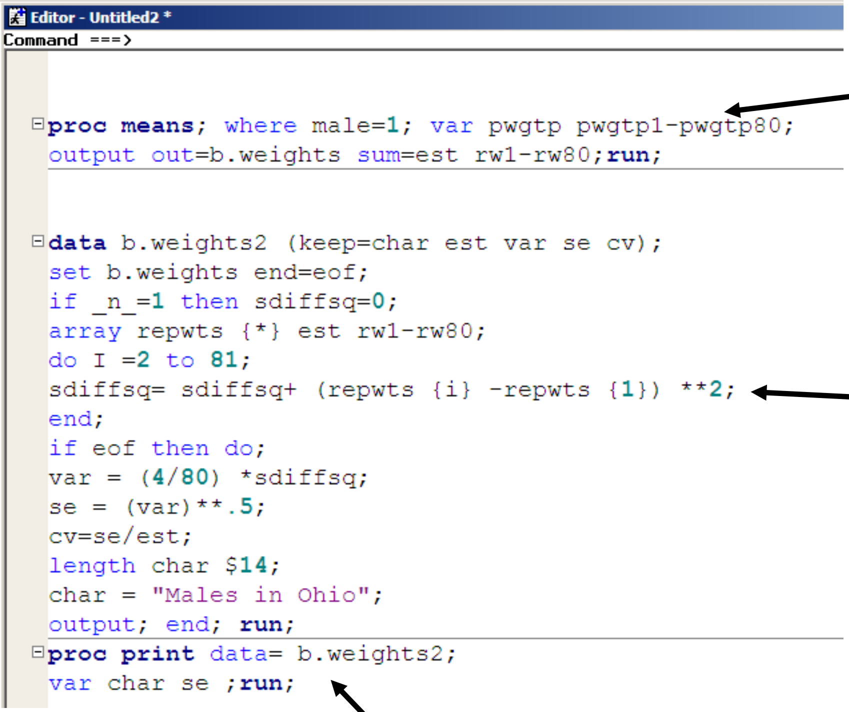
The standard error of X can be computed after the replicate estimates X_1 through X_{80} are computed. The standard error is estimated using the sum of squared differences between each replicate estimate X_r and the full sample estimate X. The standard error formula is:

$$SE(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X)^2}$$

This information comes directly from the U.S. Census Bureau's Accuracy Report. See text for link.

Using SAS:

Using the equation above, arrays are used to calculate the correct standard errors.



```
proc means; where male=1; var pwgtp pwgtp1-pwgtp80;
output out=b.weights sum=est rw1-rw80;run;

data b.weights2 (keep=char est var se cv);
set b.weights end=eof;
if _n_=1 then sdiffsq=0;
array repwts {*} est rw1-rw80;
do I =2 to 81;
sdiffsq= sdiffsq+ (repwts {i} -repwts {1}) **2;
end;
if eof then do;
var = (4/80) *sdiffsq;
se = (var)**.5;
cv=se/est;
length char $14;
char = "Males in Ohio";
output; end; run;
proc print data= b.weights2;
var char se ;run;
```

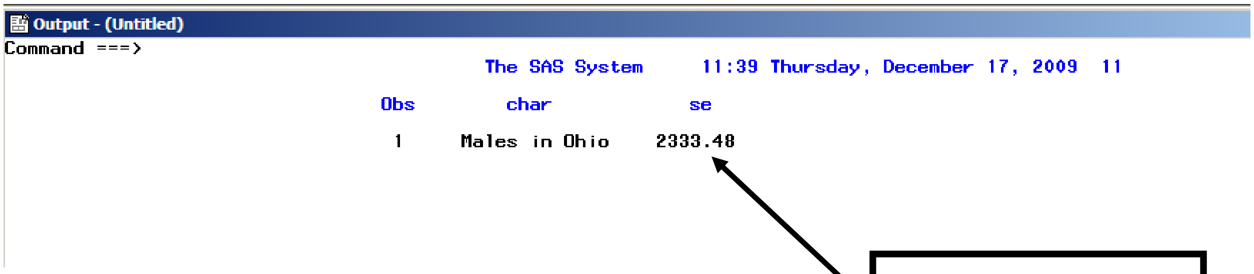
First, create a new dataset call "b.weights" that contains just the output of the sum of the person weight and all the replicate weights for males only.

Next write an array that uses this equation:

$$SE(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X)^2}$$

to calculate the correct standard errors.

Finally, include the option, se, to ensure SAS prints the standard errors.



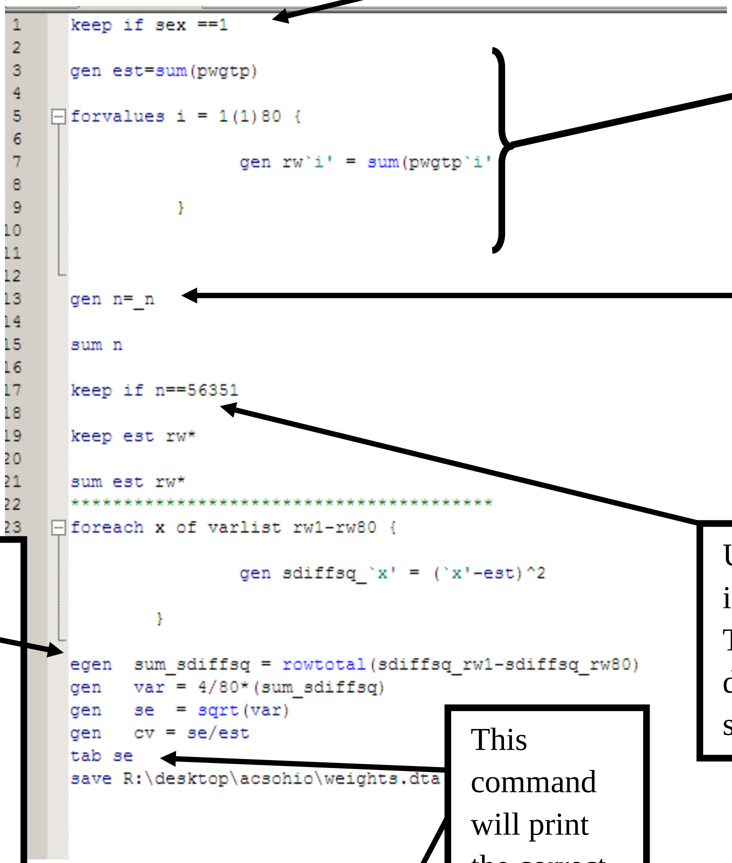
```
                          The SAS System      11:39 Thursday, December 17, 2009   11

                  Obs          char              se

                    1      Males in Ohio     2333.48
```

Here are the results and the correct standard error.

Using STATA:

Here is the do file to produce the same results as above using STATA.

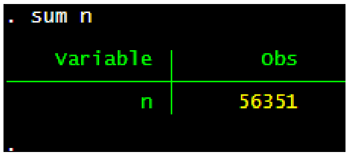
First, create a dataset with just the subsample. In this case, it is males.



```
keep if sex ==1

gen est=sum(pwgtp)

forvalues i = 1(1)80 {

                gen rw`i' = sum(pwgtp`i')

          }



gen n=_n

sum n

keep if n==56351

keep est rw*

sum est rw*
****************************************
foreach x of varlist rw1-rw80 {

                gen sdiffsq_`x' = (`x'-est)^2

          }

egen  sum_sdiffsq = rowtotal(sdiffsq_rw1-sdiffsq_rw80)
gen   var = 4/80*(sum_sdiffsq)
gen   se  = sqrt(var)
gen   cv = se/est
tab se
save R:\desktop\acsohio\weights.dta
```

Next, generate variables which are the sum of the person weight and the replicate weights.

This line of code creates an id variable. Sum the id variables. This is the output.

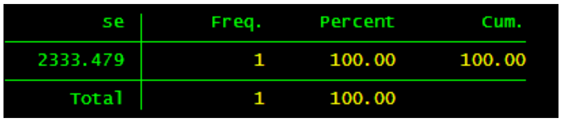


```
. sum n

    Variable |       Obs
-------------+----------
           n |     56351
.
```

Use the summed number in your keep statement. This will change depending on the subsample.

Use the above results and the egan command to calculate the correct standard error.

This command will print the correct standard error.



```
         se |      Freq.     Percent        Cum.
------------+-----------------------------------
   2333.479 |          1      100.00      100.00
------------+-----------------------------------
      Total |          1      100.00
```

Step 4: Checking Your Results

Our results from this exercise show that when using the Ohio ACS 2008 PUMS dataset, there are approximately 5,600,008 males in the state with a standard error of 2,333. The U.S. Census Bureau explains that PUMS estimates of the ACS will be slightly different than estimates from American Fact Finder because the U.S. Census Bureau does not release the full ACS dataset. The PUMS file is actually a subsample of the full dataset. As a result, analyses of ACS PUMPS data will not exactly match the values published in American Fact Finder Tables. However, the U.S. Census Bureau supplies examples of correct PUMS estimate and standard errors here: http://www.census.gov/acs/www/Products/PUMS/pumscontrols.html

We recommend trying to reproduce an estimate that the U.S. Census Bureau supplies. This strategy will assure your estimates are correct.

For further help using the American Community Survey or questions about this handout please stop by the Center for Family and Demographic Research at 5 Williams Hall or email us at CFDR@bgsu.edu.