What are some examples of commonly used survey data sets and how can they be set up for sampling purposes?

Authored by

stats writer

July 1, 2024

A survey data set is a collection of data gathered from a sample population through the use of questionnaires or surveys. Commonly used survey data sets can include demographic information, consumer behavior, or opinions and attitudes on various topics.

Some examples of commonly used survey data sets include the American Community Survey, which collects information on topics such as education, income, and housing; the National Health Interview Survey, which gathers data on health-related issues; and the General Social Survey, which collects data on social attitudes and behaviors.

When setting up a survey data set for sampling purposes, it is important to have a clear understanding of the population being studied and to use a representative sample. This can be achieved by using random sampling techniques, such as simple random sampling or stratified random sampling, to ensure that every member of the population has an equal chance of being selected for the survey. Additionally, the sample size should be large enough to accurately represent the population and reduce sampling error. Careful consideration of the survey questions and response options can also help ensure the data collected is relevant and useful for analysis. Overall, properly setting up a survey data set for sampling purposes is crucial in obtaining reliable and accurate information from a sample population.

Sample Setups for Commonly Used Survey Data Sets

This page shows the survey setups for common public use data sets in various statistical packages, including SUDAAN, Stata and SAS. If you are using an earlier version of one of these packages, the code provided below may not work. Also, please note that for your particular analysis, different sampling weight and/or replicate weights may be necessary. For data sets that contain multiple sampling weights and/or replicate weights, the documentation for

**the survey will indicate when each set of weights should be used. Many of
the setups below show the use of different weights with the same data set.
Pay special attention to this issue when merging data sets. This
page is in no way intended to be a substitute for reading the documentation for
the data set.**

**If you
would like more information on the elements of survey designs, including
sampling weights, PSUs, stratification and replicate weights, please see our page on
replicate weights.
For more information on data analysis in Stata, please see our seminar on
Survey Data Analysis
in Stata. For more information on using SUDAAN to analyze survey data,
please see our seminar**

Introduction to SUDAAN.

**A note about missing data: Many of the variables in these data sets
have special values for missing data, such as 8888 or -9. In most cases,
the statistical package (e.g., Stata, SAS, SUDAAN) will not know that these
values should be considered missing, and they will be included as legitimate
values in any analysis that is run. To convert these values to missing,
please see our
Stata FAQ if you are using Stata, and our
SAS learning module on missing data
or our SAS FAQ if you are
using either SAS or SUDAAN. Also note that the different programs handle missing data differently when you use more than one
variable in a descriptive command. For example in Stata, svy: mean x y
may give you different results than if you used two commands, svy: mean x
and svy: mean y, if x and y have different patterns of missing data.
You can (usually) quickly tell if listwise deletion is being**

used by the number of
observations being used in the analysis.

A note about non-positive probability weights or replicate weights: The different programs
handle non-positive (i.e., zero) weights differently. Stata can use cases
with non-positive sampling weights by specifying iweight instead of pweight;
hence the total number of cases read is the total number of cases used. As
a consequence, the number of raw cases used in each category in the Stata output
is different from that shown by SUDAAN or SAS. The top of the SAS output
indicates the total number of cases in the data file, as well as the number of
cases with a non-positive probability weight and the number of cases used.
The raw number of cases matches that given by SUDAAN. SUDAAN does not
count these cases as cases read in and gives a note at the top of the output.
The cases with non-positives weight are not included in

the raw frequency of cases for each category shown in the first part o the output. However, in all cases, the percent of weighted cases for each category is the same for all packages.

A note about output: We have included the output from Stata and SAS but have omitted the output from SUDAAN to save space.

This page contains the setups for the following data sets:

ACS
Add Health
CHIS
CPS
GSS
LA FANS
NHANES Continuous
NHANES III
NCS
SIPP
US Census 2000

The output from the Stata and SAS commands is shown; the output from SUDAAN has been omitted to save space.

ACS (American Community Survey)

The American Community Survey is, among other things, the replacement for the long form of the US Census. You can access one-year, three-year or five-year PUMS datasets from the ACS website. The ACS User's Guide can be found

here. The datasets (from 2005 onward) are released with successive differences replicate weights as the method of variance estimation. The documentation for this method can be found

here. For more information about the use of the replicate weights, please see http://usa.ipums.org/usa/repwt.shtml . Note that most datasets are released with both person-level and household-level

**weights (both sampling
weights and replicate weights).**

**This example uses the single year 2010 PUMS dataset, ss10hak. The
weights used are household-level weights.**

Stata

```
svyset , sdr(wgtp1 - wgtp80) vce(sdr) mse
* If negative replicate weights are a problem, specify the pweight as an iweight.
svyset , sdr(wgtp1 - wgtp80) vce(sdr) mse
svy: mean rmsp
(running mean on estimation sample)

SDR replications (80)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................  50
................................

Survey: Mean estimation Number of obs = 3,335
Population size = 307,065
Replications = 80

--------------------------------------------------------------
```

```
| SDR *
| Mean Std. Err.
-------------+------------------------------------------------
rmsp | 5.027004 .0459453 4.936953 5.117055
-------------------------------------------------------------
```

SAS

```
proc surveymeans data = acs2010 varmethod = jackknife;
weight wgtp;
repweights wgtp1 -- wgtp80 / jkcoefs = 0.05;
var rmsp;
run;

* If negative replicate weights are a problem, you might want to set them to 0;

data acs2010_nn;
set acs2010;
array temp(*) wgtp1-wgtp80;

do i = 1 to dim(temp);
if temp(i) < 0 then temp(i)=0;
end;
```

```
run;

proc surveymeans data = acs2010_nn varmethod = jackknife;
weight wgtp;
repweights wgtp1 -- wgtp80 / jkcoefs = 0.05;
var rmsp;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Observations 3335
Number of Observations Used 3071
Number of Obs with Nonpositive Weights 264
Sum of Weights 307065

Variance Estimation

Method Jackknife
Replicate Weights ACS2010_NN
Number of Replicates 80

Statistics

Std Error

Variable N Mean of Mean 95% CL for Mean

-

RMSP 3071 5.027004 0.045948 4.93556376 5.11844435

-

SUDAAN

```
proc descript data = acs2010_nn filetype = sas design =
jackknife;
weight wgtp;
jackwgts wgtp1 -- wgtp80 / adjjack = .05;
var rmsp;
setenv colwidth = 19;
setenv decwidth = 6;
run;
```

AddHealth (National Longitudinal Study of Adolescent Health, 1994-2008)

There are four waves of Add Health data. The data can be downloaded here

(University of North Carolina) or here (ICPSR). The User Guides and Documentation can be found here. There is also an excellent discussion of common mistakes to avoid when analyzing these data. Note that weight variables are in a separate dataset (one for each wave of data), so the weight variable needs to be merged with the file containing the analysis variables. Also, some of the data files contain more variables than can be read using Stata I/C (Intercooled Stata). You can use Stata S/E, Stata M/P or SAS to reduce the number of variables if you want to do your analysis in Stata I/C.

CHIS (California Health Interview Survey)

Please note that you need to register to access the CHIS data. The data and documentation can be obtained from the CHIS web site. The CHIS

methodology documentation can be found here. CHIS data are released with a sampling weight and jackknife replicate weights. The adjustment value is 1.

The 2009 adult dataset is used in the example below.

Stata

```
svyset , jkrw(rakedw1 - rakedw80, multiplier(1)) vce(jack) mse
```

pweight: rakedw0
VCE: jackknife
MSE: on
jkrweight: rakedw1 rakedw2 rakedw3 rakedw4 rakedw5 rakedw6 rakedw7 rakedw8 rakedw9 rakedw10 rakedw11 rakedw12 rakedw13
rakedw14 rakedw15 rakedw16 rakedw17 rakedw18 rakedw19 rakedw20 rakedw21 rakedw22 rakedw23 rakedw24 rakedw25
rakedw26 rakedw27 rakedw28 rakedw29 rakedw30 rakedw31 rakedw32 rakedw33 rakedw34 rakedw35 rakedw36 rakedw37
rakedw38 rakedw39 rakedw40 rakedw41 rakedw42

rakedw43 rakedw44 rakedw45 rakedw46 rakedw47 rakedw48 rakedw49
rakedw50 rakedw51 rakedw52 rakedw53 rakedw54 rakedw55 rakedw56 rakedw57 rakedw58 rakedw59 rakedw60 rakedw61
rakedw62 rakedw63 rakedw64 rakedw65 rakedw66 rakedw67 rakedw68 rakedw69 rakedw70 rakedw71 rakedw72 rakedw73
rakedw74 rakedw75 rakedw76 rakedw77 rakedw78 rakedw79 rakedw80
Strata 1: <one>
SU 1: <observations>
FPC 1: <zero>

svy: mean bmi_p
(running mean on estimation sample)

Jackknife replications (80)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..... 50
.....

Survey: Mean estimation

Number of strata = 1 Number of obs = 47614

Population size = 27546591

Replications = 80

Design df = 79

| Jknife *

| Mean Std. Err.

----------+----------

bmi_p | 26.77736 .0532953 26.67127 26.88344

SAS

```
proc surveymeans data = chis2009_adult varmethod =
jackknife;
weight rakedw0;
repweight rakedw1 -- rakedw80 / jkcoef = 1;
var bmi_p;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Observations 47614

Sum of Weights 27546591

Variance Estimation

Method Jackknife
Replicate Weights CHIS2009_ADULT
Number of Replicates 80

Statistics

Std Error
Variable N Mean of Mean 95% CL for Mean

-
BMI_P 47614 26.777356 0.053296 26.6712935 26.8834179

-

SUDAAN

```
proc descript data = chis2009_adult filetype = sas
design = jackknife;
weight rakedw0;
jackwgts rakedw1 -- rakedw80 / adjjack = 1;
var bmi_p;
```

setenv decwidth = 6;

setenv colwidth = 18;

run;

CPS (Current Population Survey)

The data and documentation can be obtained from either the
IPUMS or the
CPS website.
The CPS datasets are released with successive difference replicate weights.
For more information, please see
http://cps.ipums.org/cps/repwt.shtml . Please read the documentation
very carefully, especially with respect to how the weight variables are stored
in the dataset. You may need to divide the sampling weight by 100 and the
replicate weights by 1000 before using these weights in your analysis (depending
on where you downloaded the data).

The March 2011 supplement is used for this example.

Stata

**Note that iweight is specified instead of pweight because some of the
replicate weight values are negative. The generate and foreach
commands are provided if you need to divide your weights.**

```
gen wtsupp2 = wtsupp/100;
foreach var of varlist repwtp1 - repwtp160 {
gen n`var' = `var'/10000
}

svyset , sdrweight(repwtp1-repwtp160) vce(sdr)
svy: mean age


(running mean on estimation sample)

SDR replications (160)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................. 50
.................................................. 100
.................................................. 150
.....
```

Survey: Mean estimation Number of obs = 204983

Population size = 306109661

Replications = 160

| SDR

| Mean Std. Err.

-----------+---

age | 36.99796 .0077159 36.98284 37.01308

SAS

The data step below shows the division of weights, if it is needed.

```
data cps_3_2011;
set temp;
wtsupp2 = wtsupp/100;
array Arepwtp(160) repwtp1 - repwtp160;
array Arepwtpn(160) repwtpn1 - repwtpn160;
do x = 1 to 160;
Arepwtpn(x) = Arepwtp(x)/10000;
end;
```

```
run;

proc surveymeans data = cps_3_2011 varmethod = jackknife;
weight wtsupp;
repweights repwtp1 -- repwtp160;
var age;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Observations 204983
Sum of Weights 306109661

Variance Estimation

Method Jackknife
Replicate Weights CPS_3_2011
Number of Replicates 160

Statistics

Std Error

Variable N Mean of Mean 95% CL for Mean

-

age 204983 36.997962 0.075575 36.8487086 37.1472144

-

SUDAAN

```
proc descript data = cps_3_2011 filetype = sas design =
brr;
weight wtsupp;
repwgt repwtp1 -- repwtp160;
var age;
setenv colwidth = 19;
setenv decwidth = 3;
run;
```

GSS (General Social Survey)

The GSS data and documentation can be found here. There are datasets from 1972 to 2016.

The 2010 data are used for this example. Please note

**that although the
sampling design includes stratification, the stratification variable was not
released in the dataset.**

**NOTE: The difference in estimated population sizes between Stata and
SAS has to do with the 996 missing cases on the variable wwwhr.**

Stata

svyset sampcode

**pweight: wtssnr
VCE: linearized
Single unit: missing
Strata 1: <one>
SU 1: sampcode
FPC 1: <zero>**

**svy: mean wwwhr
(running mean on estimation sample)**

Survey: Mean estimation

Number of strata = 1 Number of obs = 1048
Number of PSUs = 79 Population size = 1084.08
Design df = 78

| Linearized
| Mean Std. Err.
---+---
wwwhr | 9.968178 .5051429 8.962516 10.97384

SAS

```
proc surveymeans data = gss2010;
weight wtssnr;
cluster sampcode;
var wwwhr;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Clusters 79
Number of Observations 2044

Sum of Weights 2043.99999

Statistics

Std Error

Variable N Mean of Mean 95% CL for Mean

-

wwwhr 1048 9.968178 0.505143 8.96251566 10.9738402

-

SUDAAN

```
proc sort data = gss2010;
by sampcode;
run;

proc descript data = gss2010 filetype = sas design = wr;
weight wtssnr;
nest sampcode;
var wwwhr;
setenv colwidth = 19;
setenv decwidth = 6;
run;
```

LA FANS (Los Angeles Family and Neighborhood Survey)

Please note that you need to register to use the L.A. FANS data. The link to the public use L.A. FANS-2 data files is here. Documentation regarding the sampling can be found here.

Stata

svyset , strata(povcat)

pweight: wgtadlt
VCE: linearized
Single unit: missing
Strata 1: povcat
SU 1: <observations>
FPC 1: <zero>

svy: mean ab5
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 3 Number of obs = 2595

Number of PSUs = 2595 Population size = 3173.95

Design df = 2592

| Linearized

| Mean Std. Err.

----------+----------

ab5 | 2.551448 .023172 2.506011 2.596886

SAS

```
proc surveymeans data = lafans1;
weight wgtadlt;
strata povcat;
var ab5;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata 3

Number of Observations 10195

Number of Observations Used 3535

Number of Obs with Nonpositive Weights 6660
Sum of Weights 3535.6271

Statistics

Std Error
Variable N Mean of Mean 95% CL for Mean

-
AB5 2595 2.551448 0.023172 2.50601087 2.59688568

-

SUDAAN

```
proc sort data = lafans1;
by povcat;
run;

proc descript data = lafans1 filetype = sas design = wr;
nest povcat;
weight wgtadlt;
var ab5;
setenv colwidth = 18;
setenv decwidth = 6;
```

run;

NHANES - Continuous (National Health and Nutrition Examination Survey)

The NHANES data and documentation can be found here. The online tutorials for these datasets are very good, and we recommend that you look these materials over before using these datasets. These tutorials also include information about combining datasets from different years. For these examples, we will use the 2009-2010 demographics dataset; the documentation for this particular dataset can be found here.
Starting in 1999, the data are released only with masked strata and PSU variables; no replicate weights are provided.

Stata

```
svyset sdmvpsu , strata(sdmvstra)
```

pweight: WTINT2YR

VCE: linearized

Single unit: missing

Strata 1: sdmvstra

SU 1: sdmvpsu

FPC 1: <zero>

svy: mean ridageyr

(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 15 Number of obs = 10537

Number of PSUs = 31 Population size = 301943719

Design df = 16

--

| Linearized

| Mean Std. Err.

-------------+--

ridageyr | 36.68331 .5459442 35.52596 37.84066

--

SAS

```
proc surveymeans data = demo_f;
cluster sdmvpsu;
strata sdmvstra;
weight wtint2yr;
var ridageyr;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata 15
Number of Clusters 31
Number of Observations 10537
Sum of Weights 301943719

Statistics

Variable	N	Mean	Std Error of Mean	95% CL for Mean	
RIDAGEYR	10537	36.683305	0.545944	35.5259552	37.8406551

-

SUDAAN

```
proc sort data = demo_f;
by sdmvstra sdmvpsu;
run;

proc descript data = demo_f filetype = sas design = wr;
weight wtint2yr;
nest sdmvstra sdmvpsu / missunit;
var ridageyr;
run;
```

NHANES III (National Health and Nutrition Examination Survey Three)

The data and documentation can obtain from the NHANES website. The NHANES III (1988 - 1994) data sets were released with the variables necessary to correct
the standard errors of the estimates by either Taylor series linearization or
the replicate weight method. To ensure the privacy of the survey

**respondents, instead of releasing the actual strata and PSU variables,
pseudo-strata and pseudo-PSU variables were released. These are used in
the same way that the "real" variables would be used. The data sets also
contain balanced-repeated replicate weights (brr). The Fay's adjustment is
1.7 or .3, depending the statistical package that you are using. Please
note that these data sets were released with multiple sampling weights and multiple sets
of replicate weights. Care must be taken to ensure that the correct
weights are being used with each analysis. The choice of weights depends
on the particular data set and variables being analyzed. In the examples
below, we use the adult data set. Note that before using the pseudo strata
and pseudo PSU variables, the data set must be sorted by the pseudo strata and
pseudo PSU. (For the data set containing only the data from 1999-2000,**

replicate weights using JK-1 are included with the data, with an adjustment of
.980769 (= 51/52). In SUDAAN, the statements would be weight wtmec2yr; jackwgts
wtmrep01 - wtmrep52 / adjjack = .980769. See guidelines.pdf for
details.)

Stata

*** with replicate weights**
*** NOTE: You need to use the formula Fay=1-1/sqrt(adjfay) to convert the value of Fay's adjustment**
given in the documentation to the form that Stata wants.
You need to use the -vce(brr)- and -mse- options
to obtain the standard errors given by SUDAAN.
display 1-(1/sqrt(1.7))
.23303501

svyset , brrweight(wtpqrp1 - wtpqrp52) fay(.23303501) vce(brr) mse

pweight: wtpfqx6
VCE: brr

MSE: on

brrweight: wtpqrp1 wtpqrp2 wtpqrp3 wtpqrp4 wtpqrp5 wtpqrp6 wtpqrp7 wtpqrp8 wtpqrp9 wtpqrp10 wtpqrp11 wtpqrp12 wtpqrp13 wtpqrp14
wtpqrp15 wtpqrp16 wtpqrp17 wtpqrp18 wtpqrp19 wtpqrp20 wtpqrp21 wtpqrp22 wtpqrp23 wtpqrp24 wtpqrp25 wtpqrp26 wtpqrp27
wtpqrp28 wtpqrp29 wtpqrp30 wtpqrp31 wtpqrp32 wtpqrp33 wtpqrp34 wtpqrp35 wtpqrp36 wtpqrp37 wtpqrp38 wtpqrp39 wtpqrp40
wtpqrp41 wtpqrp42 wtpqrp43 wtpqrp44 wtpqrp45 wtpqrp46 wtpqrp47 wtpqrp48 wtpqrp49 wtpqrp50 wtpqrp51 wtpqrp52

fay: .23303501

Strata 1: <one>

SU 1: <observations>

FPC 1: <zero>

svy: mean haznok5r

(running mean on estimation sample)

BRR replications (52)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

..... 50

..

Survey: Mean estimation Number of obs = 20014
Population size = 1.9e+08
Replications = 52
Design df = 51

| BRR *
| Mean Std. Err.
---+---
haznok5r | 6.851117 .1024657 6.645408 7.056825

---

* with pseudo-strata and pseudo-PSUs;
svyset sdppsu6 , strata(sdpstra6)

pweight: wtpfqx6
VCE: linearized
Strata 1: sdpstra6
SU 1: sdppsu6
FPC 1: <zero>

svy : mean haznok5r

(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 49 Number of obs = 20014
Number of PSUs = 98 Population size = 1.9e+08
Design df = 49

| Linearized
| Mean Std. Err.
---+---
haznok5r | 6.851117 .1237399 6.602452 7.099781

SAS

```
* with pseudo-strata and pseudo-PSUs;
proc surveymeans data = adult1;
weight wtpfqx6;
strata sdpstra6;
cluster sdppsu6;
var HAZNOK5R;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata 49
Number of Clusters 98
Number of Observations 20050
Sum of Weights 187647206

Statistics

Std Error
Variable N Mean of Mean 95% CL for Mean

-
HAZNOK5R 20014 6.851117 0.123740 6.60245228 7.09978141

-

* with brr replicate weights;
proc surveymeans data = adult1 varmethod = brr (fay = .23303501);
weight wtpfqx6;
repweights WTPQRP1 - WTPQRP52;

var HAZNOK5R;

run;

The SURVEYMEANS Procedure

Data Summary

Number of Observations 20050

Sum of Weights 187647206

Variance Estimation

Method BRR

Replicate Weights ADULT1

Number of Replicates 52

Fay Coefficient 0.23303501

Statistics

Std Error

Variable N Mean of Mean 95% CL for Mean

-

HAZNOK5R 20014 6.851117 0.102466 6.64550449 7.05672921

-

SUDAAN

```
* with brr replicate weights;
proc descript data = adult1 filetype = sas design=brr;
repwgt WTPQRP1 - WTPQRP52 / adjfay = 1.7;
weight WTPFQX6 ;
var HAZNOK5R;
setenv colwidth = 19;
setenv decwidth = 7;
print nsum wsum mean semean / nohead;
run;

* with pseudo-strata and pseudo-PSUs;
proc sort data = adult1;
by sdpstra6 sdppsu6;
run;

proc descript data = adult1 filetype = sas design = wr;
nest sdpstra6 sdppsu6 / missunit;
weight WTPFQX6 ;
var HAZNOK5R;
setenv colwidth = 19;
```

setenv decwidth = 7;

print nsum wsum mean semean / nohead;

run;

National
Comorbidity Survey

The NCS data and documentation can be obtained from the
NCS website.
(NOTE: These examples are taken from the DS2: NCS Diagnosis/Demographic
Data)

Stata

svyset secu , strata(str)

pweight: p1fwt
VCE: linearized
Strata 1: str
SU 1: secu
FPC 1: <zero>

svy: mean deplt1

(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 42 Number of obs = 8098
Number of PSUs = 84 Population size = 8098
Design df = 42

| Linearized
| Mean Std. Err.
---+---
deplt1 | .1706523 .0067263 .1570781 .1842266

SAS

A SAS SUGI 27 paper with examples from this data set can be found
here .

proc surveymeans data = ncs2;
strata str;
cluster secu;
weight p1fwt;

var deplt1;

run;

The SURVEYMEANS Procedure

Data Summary

Number of Strata 42

Number of Clusters 84

Number of Observations 8098

Sum of Weights 8097.9966

Statistics

Std Error

Variable N Mean of Mean 95% CL for Mean

-

DEPLT1 8098 0.170652 0.006726 0.15707807 0.18422659

-

SUDAAN

proc sort data = ncs2;

```
by str secu;
run;

proc descript data = ncs2 filetype = sas design = wr;
weight p1fwt;
nest str secu;
var deplt1;
setenv colwidth = 12;
setenv decwidth = 6;
print nsum wsum mean semean lowmean upmean;
run;
```

SIPP (Survey of Income and Program Participation)

The data and file setups can be downloaded from http://www.nber.org/data/sipp.html . A generic example instead of one using real data is shown below.

Stata

```
svyset , brrweight(RepWt_1-RepWt_n) fay(.5) vce(brr) mse
```

SAS

```
proc surveymeans data = sipp_data varmethod = brr fay = (.5);
weight _yourwgt;
repweights RepWt_1-RepWt_n;
var _yourvar;
run;
```

SUDAAN

```
proc descript data = sipp_data filetype = sas design=brr;
repwgt RepWt_1-RepWt_n / adjfay = 4;
weight _yourwgt;
var _yourvar;
setenv colwidth = 19;
setenv decwidth = 7;
print nsum wsum mean semean / nohead;
run;
```

US Census 2000

Census data can be obtained from the Census website.
The documentation can be found here .

**Chapter 5 describes the sampling used, and chapter 4 describes the calculations
necessary to obtain the correct standard errors (pages 4-3 to 4-15).**

**The 2000 US Census was released with person and household weights to weight the sample (either the 1% or the 5% PUMS) back to the national totals. In our examples, we
will use the person weights with person level variables. The data are
clustered within household; every person within a selected household is included
in the sample. For both institutional and non-institutional group
quarters, a pseudo household record number was assigned (see pages 2-3 and 3-1
of the documentation). Although it is clearly
stated in the documentation that the sample data set was constructed using
stratified sampling, the stratification variable was not released with the data
set. Furthermore, some of the variables used in the stratification were**

also not released, so that the stratification variable cannot be reconstructed
by the user of the data set. Hence, in our example setup, we will
ignore the stratification. Please see chapter 4 of the documentation for
instructions on how to obtain correct standard errors. In our examples, we use the 5% PUMS data for California.

Stata

Note that unless you limit the number of variables, you need to
use Stata S/E or Stata M/P.

svyset serialno

svy: tab carpool, count se cellwidth(10) format(%15.2g)
(running tabulate on estimation sample)

Number of strata = 1 Number of obs = 1690642
Number of PSUs = 616115 Population size = 33884660
Design df = 616114

----

vehicle |

occupancy | count se

----------+----------

not in u | 21347148 29827

drove al | 10418251 15947

2 people | 1572572 7244

3 people | 327968 3364

4 people | 120283 2240

5 or 6 p | 56246 1505

7 or mor | 42192 1283

|

Total | 33884660

Key: count = weighted counts

se = linearized standard errors of weighted counts

SAS

```
proc surveyfreq data = census2000;
weight pweight;
cluster serialno;
tables carpool;
```

run;

The SURVEYFREQ Procedure

Data Summary

Number of Clusters 616115
Number of Observations 1690642
Number of Observations Used 1690362
Number of Obs with Nonpositive Weights 280
Sum of Weights 33884660

vehicle occupancy

Weighted Std Dev of Std Err of
CARPOOL Frequency Frequency Wgt Freq Percent Percent

---

0 1073728 21347148 29827 62.9994 0.0452
1 511854 10418251 15947 30.7462 0.0440
2 77629 1572572 7244 4.6410 0.0207
3 16250 327968 3364 0.9679 0.0098
4 5941 120283 2240 0.3550 0.0066
5 2851 56246 1505 0.1660 0.0044
6 2109 42192 1283 0.1245 0.0038

Total 1690362 33884660 35601 100.000

SUDAAN

```
proc sort data = census2000;
by serialno;
run;

proc crosstab data = census2000 filetype = sas design = wr;
nest _one_ serialno;
weight pweight;
class carpool;setenv colwidth = 14;
print nsum wsum sewgt rowper serow;
run;
```