STATA SURVEY DATA REFERENCE MANUAL

RELEASE 12

A Stata Press Publication
StataCorp LP
College Station, Texas

Version 12

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in TEX
Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-093-9
ISBN-13: 978-1-59718-093-1

The suggested citation for this software is

StataCorp. 2011. *Stata: Release 12*. Statistical Software. College Station, TX: StataCorp LP.

Table of contents

Cross-referencing the documentation

When reading this manual, you will find references to other Stata manuals. For example,

[U] **26 Overview of Stata estimation commands**
[R] **regress**
[D] **reshape**

The first example is a reference to chapter 26, *Overview of Stata estimation commands*, in the *User's Guide*; the second is a reference to the `regress` entry in the *Base Reference Manual*; and the third is a reference to the `reshape` entry in the *Data-Management Reference Manual*.

All the manuals in the Stata Documentation have a shorthand notation:

[GSM]	*Getting Started with Stata for Mac*
[GSU]	*Getting Started with Stata for Unix*
[GSW]	*Getting Started with Stata for Windows*
[U]	*Stata User's Guide*
[R]	*Stata Base Reference Manual*
[D]	*Stata Data-Management Reference Manual*
[G]	*Stata Graphics Reference Manual*
[XT]	*Stata Longitudinal-Data/Panel-Data Reference Manual*
[MI]	*Stata Multiple-Imputation Reference Manual*
[MV]	*Stata Multivariate Statistics Reference Manual*
[P]	*Stata Programming Reference Manual*
[SEM]	*Stata Structural Equation Modeling Reference Manual*
[SVY]	*Stata Survey Data Reference Manual*
[ST]	*Stata Survival Analysis and Epidemiological Tables Reference Manual*
[TS]	*Stata Time-Series Reference Manual*
[I]	*Stata Quick Reference and Index*
[M]	*Mata Reference Manual*

Detailed information about each of these manuals may be found online at

http://www.stata-press.com/manuals/

Title

intro — Introduction to survey data manual

Description

This entry describes this manual and what has changed since Stata 11. See the next entry, [SVY] **survey**, for an introduction to Stata's survey commands.

Remarks

This manual documents the survey data commands and is referred to as [SVY] in references.

After this entry, [SVY] **survey** provides an overview of the survey commands. This manual is arranged alphabetically. If you are new to Stata's survey data commands, we recommend that you read the following sections first:

[SVY] **survey**	Introduction to survey commands
[SVY] **svyset**	Declare survey design for dataset
[SVY] **svydescribe**	Describe survey data
[SVY] **svy estimation**	Estimation commands for survey data
[SVY] **svy postestimation**	Postestimation tools for svy

Stata is continually being updated, and Stata users are continually writing new commands. To find out about the latest survey data features, type `search survey` after installing the latest official updates; see [R] **update**.

What's new

This section is intended for previous Stata users. If you are new to Stata, you may as well skip it.

1. **Survey estimation may be combined with new SEM** for structural equation modeling. See [SVY] **svy estimation** and the *Stata Structural Equation Modeling Reference Manual*.
2. **Bootstrap standard errors for survey data** using user-supplied bootstrap replicate weights. See [SVY] **svy bootstrap**.
3. **Survey SDR weights**, which is to say, successive difference replicate weights, which are supplied with many datasets from the U.S. Census Bureau. See [SVY] **svy sdr**.
4. **Contrasts**, which is to say, tests of linear hypotheses involving factor variables and their interactions from the most recently fit model. Tests include ANOVA-style tests of main effects, simple effects, interactions, and nested effects. Effects can be decomposed into comparisons with reference categories, comparisons of adjacent levels, comparisons with the grand mean, and more. New commands `contrast` and `margins, contrast` are available after `svy`; see [SVY] **svy postestimation**. Also see [R] **contrast** and [R] **margins, contrast**.
5. **Pairwise comparisons** of means, estimated cell means, estimated marginal means, predictive margins of linear and nonlinear responses, intercepts, and slopes. In addition to ANOVA-style comparisons, comparisons can be made of population averages. New commands `pwcompare` and `margins, pwcompare` are available after `svy`; see [SVY] **svy postestimation**. Also see [R] **pwcompare** and [R] **margins, pwcompare**.

6. **Graphs of margins, marginal effects, contrasts, and pairwise comparisons**. Margins and effects can be obtained from linear or nonlinear (for example, probability) responses. New command `marginsplot` is available after `svy`; see [SVY] **svy postestimation**. Also see [R] **marginsplot**.
7. **Estimation output improved.**
   a. **Implied zero coefficients now shown**. When a coefficient is omitted, it is now shown as being zero and the reason it was omitted—collinearity, base, empty—is shown in the standard-error column. (The word "omitted" is shown if the coefficient was omitted because of collinearity.)
   b. **You can set displayed precision for all values in coefficient tables** using `set cformat`, `set pformat`, and `set sformat`. Or you may use options `cformat()`, `pformat()`, and `sformat()` now allowed on all estimation commands. See [R] **set cformat** and [R] **estimation options**.
   c. **Estimation commands now respect the width of the Results window.** This feature may be turned off by new display option `nolstretch`. See [R] **estimation options**.
   d. **You can now set whether base levels, empty cells, and omitted are shown** using `set showbaselevels`, `set showemptycells`, and `set showomitted`. See [R] **set showbaselevels**.
8. **Survey goodness-of-fit** available after `logistic`, `logit`, and `probit` with new command `estat gof`. See [SVY] **estat**.
9. **Survey coefficient of variation (CV)** available with new command `estat cv`. See [SVY] **estat**.

For a list of all the new features in Stata 12, see [U] **1.3 What's new**.

Also see

[U] **1.3 What's new**

[R] **intro** — Introduction to base reference manual

Title

survey — Introduction to survey commands

Description

The *Survey Data Reference Manual* is organized alphabetically, making it easy to find an individual entry if you know the name of a command. This overview organizes and presents the commands conceptually, that is, according to the similarities in the functions they perform.

Survey design tools

[SVY] **svyset**	Declare survey design for dataset
[SVY] **svydescribe**	Describe survey data

Survey data analysis tools

[SVY] **svy**	The survey prefix command
[SVY] **svy estimation**	Estimation commands for survey data
[SVY] **svy: tabulate oneway**	One-way tables for survey data
[SVY] **svy: tabulate twoway**	Two-way tables for survey data
[SVY] **svy postestimation**	Postestimation tools for svy
[SVY] **estat**	Postestimation statistics for survey data, such as design effects
[SVY] **svy bootstrap**	Bootstrap for survey data
[SVY] ***bootstrap_options***	More options for bootstrap variance estimation
[SVY] **svy brr**	Balanced repeated replication for survey data
[SVY] ***brr_options***	More options for BRR variance estimation
[SVY] **svy jackknife**	Jackknife estimation for survey data
[SVY] ***jackknife_options***	More options for jackknife variance estimation
[SVY] **svy sdr**	Successive difference replication for survey data
[SVY] ***sdr_options***	More options for SDR variance estimation

Survey data concepts

[SVY] **variance estimation**	Variance estimation for survey data
[SVY] **subpopulation estimation**	Subpopulation estimation for survey data
[SVY] **direct standardization**	Direct standardization of means, proportions, and ratios
[SVY] **poststratification**	Poststratification for survey data

Tools for programmers of new survey commands

[SVY] **ml for svy**	Maximum pseudolikelihood estimation for survey data
[SVY] **svymarkout**	Mark observations for exclusion on the basis of survey characteristics

Remarks

Remarks are presented under the following headings:

Introduction

Stata's facilities for survey data analysis are centered around the `svy` prefix command. After you identify the survey design characteristics with the `svyset` command, prefix the estimation commands in your data analysis with "`svy:`". For example, where you would normally use the `regress` command to fit a linear regression model for nonsurvey data, use `svy: regress` to fit a linear regression model for your survey data.

Why should you use the `svy` prefix command when you have survey data? To answer this question, we need to discuss some of the characteristics of survey design and survey data collection because these characteristics affect how we must perform our analysis if we want to get it right.

Survey data are characterized by the following:

- Sampling weights, also called probability weights—`pweights` in Stata's terminology
- Cluster sampling
- Stratification

These features arise from the design and details of the data collection procedure. Here's a brief description of how these design features affect the analysis of the data:

- *Sampling weights.* In sample surveys, observations are selected through a random process, but different observations may have different probabilities of selection. Weights are equal to (or proportional to) the inverse of the probability of being sampled. Various postsampling adjustments to the weights are sometimes made, as well. A weight of w_j for the jth observation means, roughly speaking, that the jth observation represents w_j elements in the population from which the sample was drawn.

  Omitting weights from the analysis results in estimates that may be biased, sometimes seriously so. Sampling weights also play a role in estimating standard errors.

- *Clustering.* Individuals are not sampled independently in most survey designs. Collections of individuals (for example, counties, city blocks, or households) are typically sampled as a group, known as a *cluster*.

  There may also be further subsampling within the clusters. For example, counties may be sampled, then city blocks within counties, then households within city blocks, and then finally persons within households. The clusters at the first level of sampling are called *primary sampling units* (PSUs)—in this example, counties are the PSUs. In the absence of clustering, the PSUs are defined to be the individuals, or, equivalently, clusters, each of size one.

  Cluster sampling typically results in larger sample-to-sample variability than sampling individuals directly. This increased variability must be accounted for in standard error estimates, hypothesis testing, and other forms of inference.

- *Stratification.* In surveys, different groups of clusters are often sampled separately. These groups are called *strata*. For example, the 254 counties of a state might be divided into two strata, say, urban counties and rural counties. Then 10 counties might be sampled from the urban stratum, and 15 from the rural stratum.

  Sampling is done independently across strata; the stratum divisions are fixed in advance. Thus strata are statistically independent and can be analyzed as such. When the individual strata are more homogeneous than the population as a whole, the homogeneity can be exploited to produce smaller (and honestly so) estimates of standard errors.

To put it succinctly: using sampling weights is important to get the point estimates right. We must consider the weighting, clustering, and stratification of the survey design to get the standard errors right. If our analysis ignores the clustering in our design, we would probably produce standard errors that are smaller than they should be. Stratification can be used to get smaller standard errors for a given overall sample size.

For more detailed introductions to complex survey data analysis, see Cochran (1977); Heeringa, West, and Berglund (2010); Kish (1965); Levy and Lemeshow (2008); Scheaffer et al.; (2012); Skinner, Holt, and Smith (1989); Stuart (1984); Thompson (2002); and Williams (1978).

Survey design tools

Before using `svy`, first take a quick look at [SVY] **svyset**. Use the `svyset` command to specify the variables that identify the survey design characteristics and default method for estimating standard errors. Once set, `svy` will automatically use these design specifications until they are cleared or changed or a new dataset is loaded into memory.

As the following two examples illustrate, `svyset` allows you to identify a wide range of complex sampling designs. First, we show a simple single-stage design and then a complex multistage design.

▷ Example 1: Survey data from a one-stage design

A commonly used single-stage survey design uses clustered sampling across several strata, where the clusters are sampled without replacement. In a Stata dataset composed of survey data from this design, the survey design variables identify information about the strata, PSUs (clusters), sampling weights, and finite population correction. Here we use `svyset` to specify these variables, respectively named `strata`, `su1`, `pw`, and `fpc1`.

```
. use http://www.stata-press.com/data/r12/stage5a
. svyset su1 [pweight=pw], strata(strata) fpc(fpc1)
      pweight: pw
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: su1
        FPC 1: fpc1
```

In addition to the variables we specified, `svyset` reports that the default method for estimating standard errors is Taylor linearization and that `svy` will report missing values for the standard errors when it encounters a stratum with one sampling unit (also called singleton strata).

◁

▷ Example 2: Multistage survey data

We have (fictional) data on American high school seniors (12th graders), and the data were collected according to the following multistage design. In the first stage, counties were independently selected within each state. In the second stage, schools were selected within each chosen county. Within each chosen school, a questionnaire was filled out by every attending high school senior. We have entered all the information into a Stata dataset called `multistage.dta`.

The survey design variables are as follows:

- `state` contains the stratum identifiers.
- `county` contains the first-stage sampling units.
- `ncounties` contains the total number of counties within each state.
- `school` contains the second-stage sampling units.
- `nschools` contains the total number of schools within each county.
- `sampwgt` contains the sampling weight for each sampled individual.

Here we load the dataset into memory and use `svyset` with the above variables to declare that these data are survey data.

```
. use http://www.stata-press.com/data/r12/multistage
. svyset county [pw=sampwgt], strata(state) fpc(ncounties) || school, fpc(nschools)
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
. save highschool
file highschool.dta saved
```

We saved the `svyset` dataset to `highschool.dta`. We can now `use` this new dataset without having to worry about respecifying the design characteristics.

```
. clear
. describe
Contains data
  obs:             0
 vars:             0
 size:             0
Sorted by:
. use highschool
. svyset
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

◁

After the design characteristics have been `svyset`, you should also look at [SVY] **svydescribe**. Use `svydescribe` to browse each stage of your survey data; `svydescribe` reports useful information on sampling unit counts, missing data, and singleton strata.

▷ Example 3: Survey describe

Here we use `svydescribe` to describe the first stage of our survey dataset of sampled high school seniors. We specified the `weight` variable to get `svydescribe` to report on where it contains missing values and how this affects the estimation sample.

```
. svydescribe weight
Survey: Describing stage 1 sampling units
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

Stratum	#Units included	#Units omitted	#Obs with complete data	#Obs with missing data	#Obs per included Unit min	mean	max
1	2	0	92	0	34	46.0	58
2	2	0	112	0	51	56.0	61
3	2	0	43	0	18	21.5	25
4	2	0	37	0	14	18.5	23
5	2	0	96	0	38	48.0	58
(*output omitted*)							
46	2	0	115	0	56	57.5	59
47	2	0	67	0	28	33.5	39
48	2	0	56	0	23	28.0	33
49	2	0	78	0	39	39.0	39
50	2	0	64	0	31	32.0	33
50	100	0	4071	0	14	40.7	81
			4071				

From the output, we gather that there are 50 strata, each stratum contains two PSUs, the PSUs vary in size, and the total sample size is 4,071 students. We can also see that there are no missing data in the `weight` variable.

◁

Survey data analysis tools

Stata's suite of survey data commands is governed by the `svy` prefix command; see [SVY] **svy** and [SVY] **svy estimation**. `svy` runs the supplied estimation command while accounting for the survey design characteristics in the point estimates and variance estimation method. The available variance estimation methods are balanced repeated replication (BRR), the bootstrap, the jackknife, successive difference replication, and first-order Taylor linearization. By default, `svy` computes standard errors by using the linearized variance estimator—so called because it is based on a first-order Taylor series linear approximation (Wolter 2007). In the nonsurvey context, we refer to this variance estimator as the *robust* variance estimator, otherwise known in Stata as the Huber/White/sandwich estimator; see [P] **_robust**.

▷ Example 4: Estimating a population mean

Here we use the `svy` prefix with the `mean` command to estimate the average weight of high school seniors in our population.

```
. svy: mean weight
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      50          Number of obs    =     4071
Number of PSUs   =     100          Population size  =  8000000
                                    Design df        =       50

--------------------------------------------------------------
             |             Linearized
             |        Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      weight |    160.2863   .7412512      158.7974    161.7751
--------------------------------------------------------------
```

In its header, `svy` reports the number of strata and PSUs from the first stage, the sample size, an estimate of population size, and the design degrees of freedom. Just like the standard output from the `mean` command, the table of estimation results contains the estimated mean and its standard error as well as a confidence interval.

◁

▷ Example 5: Survey regression

Here we use the `svy` prefix with the `regress` command to model the association between `weight` and `height` in our population of high school seniors.

```
. svy: regress weight height
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        50                  Number of obs      =      4071
Number of PSUs     =       100                  Population size    =   8000000
                                                Design df          =        50
                                                F(   1,     50)    =    593.99
                                                Prob > F           =    0.0000
                                                R-squared          =    0.2787

------------------------------------------------------------------------------
             |             Linearized
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .7163115   .0293908    24.37   0.000     .6572784    .7753447
       _cons |  -149.6183   12.57265   -11.90   0.000    -174.8712   -124.3654
------------------------------------------------------------------------------
```

In addition to the header elements we saw in the previous example using `svy: mean`, the command `svy: regress` also reports a model F test and estimated R^2. Although many of Stata's model-fitting commands report Z statistics for testing coefficients against zero, `svy` always reports t statistics and uses the design degrees of freedom to compute p-values.

◁

The svy prefix can be used with many estimation commands in Stata. Here is the list of estimation commands that support the svy prefix.

Descriptive statistics

mean	[R] **mean** — Estimate means
proportion	[R] **proportion** — Estimate proportions
ratio	[R] **ratio** — Estimate ratios
total	[R] **total** — Estimate totals

Linear regression models

cnsreg	[R] **cnsreg** — Constrained linear regression
glm	[R] **glm** — Generalized linear models
intreg	[R] **intreg** — Interval regression
nl	[R] **nl** — Nonlinear least-squares estimation
regress	[R] **regress** — Linear regression
tobit	[R] **tobit** — Tobit regression
treatreg	[R] **treatreg** — Treatment-effects model
truncreg	[R] **truncreg** — Truncated regression

Structural equation models

[SEM] *Stata Structural Equation Modeling Reference Manual*

Survival-data regression models

stcox	[ST] **stcox** — Cox proportional hazards model
streg	[ST] **streg** — Parametric survival models

Binary-response regression models

biprobit	[R] **biprobit** — Bivariate probit regression
cloglog	[R] **cloglog** — Complementary log-log regression
hetprob	[R] **hetprob** — Heteroskedastic probit model
logistic	[R] **logistic** — Logistic regression, reporting odds ratios
logit	[R] **logit** — Logistic regression, reporting coefficients
probit	[R] **probit** — Probit regression
scobit	[R] **scobit** — Skewed logistic regression

Discrete-response regression models

clogit	[R] **clogit** — Conditional (fixed-effects) logistic regression
mlogit	[R] **mlogit** — Multinomial (polytomous) logistic regression
mprobit	[R] **mprobit** — Multinomial probit regression
ologit	[R] **ologit** — Ordered logistic regression
oprobit	[R] **oprobit** — Ordered probit regression
slogit	[R] **slogit** — Stereotype logistic regression

Poisson regression models

`gnbreg`	Generalized negative binomial regression in [R] **nbreg**
`nbreg`	[R] **nbreg** — Negative binomial regression
`poisson`	[R] **poisson** — Poisson regression
`tnbreg`	[R] **tnbreg** — Truncated negative binomial regression
`tpoisson`	[R] **tpoisson** — Truncated Poisson regression
`zinb`	[R] **zinb** — Zero-inflated negative binomial regression
`zip`	[R] **zip** — Zero-inflated Poisson regression

Instrumental-variables regression models

`ivprobit`	[R] **ivprobit** — Probit model with continuous endogenous regressors
`ivregress`	[R] **ivregress** — Single-equation instrumental-variables regression
`ivtobit`	[R] **ivtobit** — Tobit model with continuous endogenous regressors

Regression models with selection

`heckman`	[R] **heckman** — Heckman selection model
`heckprob`	[R] **heckprob** — Probit model with sample selection

Stata is continually being updated. To view the most up-to-date list of estimation commands that support the `svy` prefix, type `help svy estimation`.

▷ Example 6: Cox's proportional hazards model

Suppose that we want to model the incidence of lung cancer by using three risk factors: smoking status, sex, and place of residence. Our dataset comes from a longitudinal health survey: the First National Health and Nutrition Examination Survey (NHANES I) (Miller 1973; Engel et al. 1978) and its 1992 Epidemiologic Follow-up Study (NHEFS) (Cox et al. 1997); see the National Center for Health Statistics website at http://www.cdc.gov/nchs/. We will be using data from the samples identified by NHANES I examination locations 1–65 and 66–100; thus we will `svyset` the revised pseudo-PSU and strata variables associated with these locations. Similarly, our `pweight` variable was generated using the sampling weights for the nutrition and detailed samples for locations 1–65 and the weights for the detailed sample for locations 66–100.

```
. use http://www.stata-press.com/data/r12/nhefs
. svyset psu2 [pw=swgt2], strata(strata2)
      pweight: swgt2
          VCE: linearized
  Single unit: missing
     Strata 1: strata2
         SU 1: psu2
        FPC 1: <zero>
```

The lung cancer information was taken from the 1992 NHEFS interview data. We use the participants' ages for the time scale. Participants who never had lung cancer and were alive for the 1992 interview were considered censored. Participants who never had lung cancer and died before the 1992 interview were also considered censored at their age of death.

```
. stset age_lung_cancer [pw=swgt2], fail(lung_cancer)
     failure event:  lung_cancer != 0 & lung_cancer < .
obs. time interval:  (0, age_lung_cancer]
 exit on or before:  failure
           weight:  [pweight=swgt2]
--------------------------------------------------------------------------
    14407  total obs.
     5126  event time missing (age_lung_cancer>=.)             PROBABLE ERROR
--------------------------------------------------------------------------
     9281  obs. remaining, representing
       83  failures in single record/single failure data
   599691  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        97
```

Although `stset` warns us that it is a "probable error" to have 5,126 observations with missing event times, we can verify from the 1992 NHEFS documentation that there were indeed 9,281 participants with complete information.

For our proportional hazards model, we pulled the risk factor information from the NHANES I and 1992 NHEFS datasets. Smoking status was taken from the 1992 NHEFS interview data, but we filled in all but 132 missing values by using the general medical history supplement data in NHANES I. Smoking status is represented by separate indicator variables for former smokers and current smokers; the base comparison group is nonsmokers. Sex was determined using the 1992 NHEFS vitality data and is represented by an indicator variable for males. Place-of-residence information was taken from the medical history questionnaire in NHANES I and is represented by separate indicator variables for rural and heavily populated (more than 1 million people) urban residences; the base comparison group is urban residences with populations of fewer than 1 million people.

```
. svy: stcox former_smoker smoker male urban1 rural
(running stcox on estimation sample)
Survey: Cox regression
Number of strata   =        35                  Number of obs      =       9149
Number of PSUs     =       105                  Population size    =  151327827
                                                Design df          =         70
                                                F(   5,     66)    =      14.07
                                                Prob > F           =     0.0000
```

_t	Haz. Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
former_smo~r	2.788113	.6205102	4.61	0.000	1.788705	4.345923
smoker	7.849483	2.593249	6.24	0.000	4.061457	15.17051
male	1.187611	.3445315	0.59	0.555	.6658757	2.118142
urban1	.8035074	.3285144	-0.54	0.594	.3555123	1.816039
rural	1.581674	.5281859	1.37	0.174	.8125799	3.078702

From the above results, we can see that both former and current smokers have a significantly higher risk for developing lung cancer than do nonsmokers.

◁

`svy: tabulate` can be used to produce one-way and two-way tables with survey data and can produce survey-adjusted tests of independence for two-way contingency tables; see [SVY] **svy: tabulate oneway** and [SVY] **svy: tabulate twoway**.

Example 7: Two-way tables for survey data

With data from the Second National Health and Nutrition Examination Survey (NHANES II) (McDowell et al. 1981), we use `svy: tabulate` to produce a two-way table of cell proportions along with their standard errors and confidence intervals (the survey design characteristics have already been `svyset`). We also use the `format()` option to get `svy: tabulate` to report the cell values and marginals to four decimal places.

```
. use http://www.stata-press.com/data/r12/nhanes2b
. svy: tabulate race diabetes, row se ci format(%7.4f)
(running tabulate on estimation sample)
Number of strata   =        31                  Number of obs      =      10349
Number of PSUs     =        62                  Population size    =  117131111
                                                Design df          =         31
```

1=white, 2=black, 3=other	diabetes, 1=yes, 0=no 0	1	Total
White	0.9680	0.0320	1.0000
	(0.0020)	(0.0020)	
	[0.9638,0.9718]	[0.0282,0.0362]	
Black	0.9410	0.0590	1.0000
	(0.0061)	(0.0061)	
	[0.9271,0.9523]	[0.0477,0.0729]	
Other	0.9797	0.0203	1.0000
	(0.0076)	(0.0076)	
	[0.9566,0.9906]	[0.0094,0.0434]	
Total	0.9658	0.0342	1.0000
	(0.0018)	(0.0018)	
	[0.9619,0.9693]	[0.0307,0.0381]	

```
  Key:  row proportions
        (linearized standard errors of row proportions)
        [95% confidence intervals for row proportions]
  Pearson:
    Uncorrected   chi2(2)         =   21.3483
    Design-based  F(1.52, 47.26)  =   15.0056     P = 0.0000
```

`svy: tabulate` has many options, such as the `format()` option, for controlling how the table looks. See [SVY] **svy: tabulate twoway** for a discussion of the different design-based and unadjusted tests of association.

◁

All the standard postestimation commands (for example, `estimates`, `lincom`, `margins`, `nlcom`, `test`, `testnl`) are also available after `svy`.

▷ Example 8: Comparing means

Going back to our high school survey data in example 2, we estimate the mean of `weight` (in pounds) for each subpopulation identified by the categories of the `sex` variable (male and female).

```
. use http://www.stata-press.com/data/r12/highschool
. svy: mean weight, over(sex)
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      50          Number of obs    =     4071
Number of PSUs   =     100          Population size  =  8000000
                                    Design df        =       50
          male: sex = male
        female: sex = female
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
weight				
male	175.4809	1.116802	173.2377	177.7241
female	146.204	.9004157	144.3955	148.0125

Here we use the `test` command to test the hypothesis that the average male is 30 pounds heavier than the average female; from the results, we cannot reject this hypothesis at the 5% level.

```
. test [weight]male - [weight]female = 30
Adjusted Wald test
 ( 1)  [weight]male - [weight]female = 30
       F(  1,    50) =    0.23
            Prob > F =    0.6353
```

◁

`estat` has specific subroutines for use after `svy`; see [SVY] **estat**.

- `estat svyset` reports the survey design settings used to produce the current estimation results.
- `estat effects` and `estat lceffects` report a table of design and misspecification effects for point estimates and linear combinations of point estimates, respectively.
- `estat size` reports a table of sample and subpopulation sizes after `svy: mean`, `svy: proportion`, `svy: ratio`, and `svy: total`.
- `estat sd` reports subpopulation standard deviations on the basis of the estimation results from `mean` and `svy: mean`.
- `estat strata` reports the number of singleton and certainty strata within each sampling stage.
- `estat cv` reports the coefficient of variation for each coefficient in the current estimation results.
- `estat gof` reports a goodness-of-fit test for binary response models using survey data.

▷ Example 9: Design effects

Here we use `estat effects` to report the design effects DEFF and DEFT for the mean estimates from the previous example.

```
. estat effects
         male: sex = male
       female: sex = female
```

Over	Mean	Linearized Std. Err.	DEFF	DEFT
weight				
male	175.4809	1.116802	2.61016	1.61519
female	146.204	.9004157	1.7328	1.31603

```
Note: weights must represent population totals for deff to
      be correct when using an FPC; however, deft is
      invariant to the scale of weights.
```

Now we use `estat lceffects` to report the design effects DEFF and DEFT for the difference of the mean estimates from the previous example.

```
. estat lceffects [weight]male - [weight]female
 ( 1)  [weight]male - [weight]female = 0
```

Mean	Coef.	Std. Err.	DEFF	DEFT
(1)	29.27691	1.515201	2.42759	1.55768

```
Note: weights must represent population totals for deff to
      be correct when using an FPC; however, deft is
      invariant to the scale of weights.
```

◁

The `svy brr` prefix command produces point and variance estimates by using the BRR method; see [SVY] **svy brr**. BRR was first introduced by McCarthy (1966, 1969a, and 1969b) as a method of variance estimation for designs with two PSUs in every stratum. The BRR variance estimator tends to give more reasonable variance estimates for this design than the linearized variance estimator, which can result in large values and undesirably wide confidence intervals.

The `svy jackknife` prefix command produces point and variance estimates by using the jackknife replication method; see [SVY] **svy jackknife**. The jackknife is a data-driven variance estimation method that can be used with model-fitting procedures for which the linearized variance estimator is not implemented, even though a linearized variance estimator is theoretically possible to derive (Shao and Tu 1995).

To protect the privacy of survey participants, public survey datasets may contain replicate-weight variables instead of variables that identify the PSUs and strata. These replicate-weight variables can be used with the appropriate replication method for variance estimation instead of the linearized variance estimator; see [SVY] **svyset**.

The `svy brr` and `svy jackknife` prefix commands can be used with those commands that may not be fully supported by `svy` but are compatible with the BRR and the jackknife replication methods. They can also be used to produce point estimates for expressions of estimation results from a prefixed command.

The `svy bootstrap` and `svy sdr` prefix commands work only with replicate weights. Both assume that you have obtained these weight variables externally.