Title

stata.com

svydescribe — Describe survey data

Description

`svydescribe` displays a table that describes the strata and the sampling units for a given sampling stage in a survey dataset.

Menu

Statistics > Survey data analysis > Setup and utilities > Describe survey data

Syntax

`svydescribe` [*varlist*] [*if*] [*in*] [, *options*]

options	Description
Main	
`stage(#)`	sampling stage to describe; default is `stage(1)`
`finalstage`	display information per sampling unit in the final stage
`single`	display only the strata with one sampling unit
`generate(`*newvar*`)`	generate a variable identifying strata with one sampling unit

`svydescribe` requires that the survey design variables be identified using `svyset`; see [SVY] **svyset**.

Options

Main

`stage(#)` specifies the sampling stage to describe. The default is `stage(1)`.

`finalstage` specifies that results be displayed for each sampling unit in the final sampling stage; that is, a separate line of output is produced for every sampling unit in the final sampling stage. This option is not allowed with `stage()`, `single`, or `generate()`.

`single` specifies that only the strata containing one sampling unit be displayed in the table.

`generate(`*newvar*`)` stores a variable that identifies strata containing one sampling unit for a given sampling stage.

Remarks and examples

stata.com

Survey datasets are typically the result of a stratified survey design with cluster sampling in one or more stages. Within a stratum for a given sampling stage, there are sampling units, which may be either clusters of observations or individual observations.

`svydescribe` displays a table that describes the strata and sampling units for a given sampling stage. One row of the table is produced for each stratum. Each row contains the number of sampling units, the range and mean of the number of observations per sampling unit, and the total number of observations. If the `finalstage` option is specified, one row of the table is produced for each sampling unit of the final stage. Here each row contains the number of observations for the respective sampling unit.

If a varlist is specified, `svydescribe` reports the number of sampling units that contain at least one observation with complete data (that is, no missing values) for all variables in *varlist*. These are the sampling units that would be used to compute point estimates by using the variables in *varlist* with a given `svy` estimation command.

▷ Example 1: Strata with one sampling unit

We use data from the Second National Health and Nutrition Examination Survey (NHANES II) (McDowell et al. 1981) as our example. First, we set the PSU, `pweight`, and strata variables.

```
. use http://www.stata-press.com/data/r15/nhanes2b
. svyset psuid [pweight=finalwgt], strata(stratid)
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

svydescribe will display the strata and PSU arrangement of the dataset.

```
. svydescribe
Survey: Describing stage 1 sampling units
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	2	380	165	190.0	215
2	2	185	67	92.5	118
3	2	348	149	174.0	199
(*output omitted*)					
17	2	393	180	196.5	213
18	2	359	144	179.5	215
20	2	285	125	142.5	160
21	2	214	102	107.0	112
(*output omitted*)					
31	2	308	143	154.0	165
32	2	450	211	225.0	239
31	62	10,351	67	167.0	288

Our NHANES II dataset has 31 strata (stratum 19 is missing) and two PSUs per stratum.

The hdresult variable contains serum levels of high-density lipoprotein (HDL). If we try to estimate the mean of hdresult, we get a missing value for the standard-error estimate and a note explaining why.

```
. svy: mean hdresult
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size  =  98,725,345
                                  Design df        =          29
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
hdresult	49.67141	.	.	.

```
Note: Missing standard error because of stratum with single
      sampling unit.
```

Running svydescribe with hdresult and the single option will show which strata have only one PSU.

```
. svydescribe hdresult, single

Survey: Describing strata with a single sampling unit in stage 1

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

Stratum	#Units included	#Units omitted	#Obs with complete data	#Obs with missing data	#Obs per included Unit min	mean	max
1	1*	1	114	266	114	114.0	114
2	1*	1	98	87	98	98.0	98
	2						

Both stratid = 1 and stratid = 2 have only one PSU with nonmissing values of hdresult. Because this dataset has only 62 PSUs, the finalstage option produces a manageable amount of output:

```
. svydescribe hdresult, finalstage

Survey: Describing final stage sampling units

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

Stratum	Unit	#Obs with complete data	#Obs with missing data
1	1	0	215
1	2	114	51
2	1	98	20
2	2	0	67
(*output omitted*)			
32	2	203	8
31	62	8,720	1,631
			10,351

It is rather striking that there are two PSUs with no values for hdresult. All other PSUs have only a moderate number of missing values. Obviously, here a data analyst should first try to ascertain why these data are missing. The answer here (C. L. Johnson, 1995, pers. comm.) is that HDL measurements could not be collected until the third survey location. Thus there are no hdresult data for the first two locations: stratid = 1, psuid = 1 and stratid = 2, psuid = 2.

Assuming that we wish to go ahead and analyze the hdresult data, we must collapse strata—that is, merge them—so that every stratum has at least two PSUs with some nonmissing values. We can accomplish this by collapsing stratid = 1 into stratid = 2. To perform the stratum collapse, we create a new strata identifier, newstr, and a new PSU identifier, newpsu.

```
. generate newstr = stratid
. generate newpsu = psuid
. replace newpsu = psuid + 2 if stratid == 1
(380 real changes made)
. replace newstr = 2 if stratid == 1
(380 real changes made)
```

`svyset` the new PSU and strata variables.

```
. svyset newpsu [pweight=finalwgt], strata(newstr)
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: newstr
         SU 1: newpsu
        FPC 1: <zero>
```

Then use `svydescribe` to check what we have done.

```
. svydescribe hdresult, finalstage
Survey: Describing final stage sampling units
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: newstr
         SU 1: newpsu
        FPC 1: <zero>
```

Stratum	Unit	#Obs with complete data	#Obs with missing data
2	1	98	20
2	2	0	67
2	3	0	215
2	4	114	51
3	1	161	38
3	2	116	33
(*output omitted*)			
32	1	180	59
32	2	203	8
30	62	8,720	1,631
		10,351	

The new stratum, `newstr` = 2, has four PSUs, two of which contain some nonmissing values of `hdresult`. This is sufficient to allow us to estimate the mean of `hdresult` and get a nonmissing standard-error estimate.

```
. svy: mean hdresult
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      30       Number of obs    =       8,720
Number of PSUs   =      60       Population size =  98,725,345
                                 Design df       =          30
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
hdresult	49.67141	.3830147	48.88919	50.45364

Example 2: Using e(sample) to find strata with one sampling unit

Some estimation commands drop observations from the estimation sample when they encounter collinear predictors or perfect predictors. Ascertaining which strata contain one sampling unit is therefore difficult. We can then use `if e(sample)` instead of *varlist* when faced with the problem of strata with one sampling unit. We revisit the previous analysis to illustrate.

```
. use http://www.stata-press.com/data/r15/nhanes2b, clear
. svy: mean hdresult
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size =  98,725,345
                                 Design df       =          29
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
hdresult	49.67141	.	.	.

```
Note: Missing standard error because of stratum with single
      sampling unit.
. svydescribe if e(sample), single
Survey: Describing strata with a single sampling unit in stage 1
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

Stratum	#Units	#Obs	#Obs per Unit min	mean	max
1	1*	114	114	114.0	114
2	1*	98	98	98.0	98
	2				

Methods and formulas

See Eltinge and Sribney (1996) for an earlier implementation of `svydescribe`.

References

Eltinge, J. L., and W. M. Sribney. 1996. svy3: Describing survey data: Sampling design and missing data. *Stata Technical Bulletin* 31: 23–26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 235–239. College Station, TX: Stata Press.

McDowell, A., A. Engel, J. T. Massey, and K. Maurer. 1981. Plan and operation of the Second National Health and Nutrition Examination Survey, 1976–1980. *Vital and Health Statistics* 1(15): 1–144.

Also see

[SVY] **svy** — The survey prefix command

[SVY] **svyset** — Declare survey design for dataset

[SVY] **survey** — Introduction to survey commands

[SVY] **variance estimation** — Variance estimation for survey data