svydescribe — 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.

Menu

Statistics > Survey data analysis > Setup and utilities > 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.

Options

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.
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.

```stata
use http://www.stata-press.com/data/r13/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.

```stata
svydescribe
```

<table>
<thead>
<tr>
<th>Stratum</th>
<th>#Units</th>
<th>#Obs</th>
<th>min</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>380</td>
<td>165</td>
<td>190.0</td>
<td>215</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>185</td>
<td>67</td>
<td>92.5</td>
<td>118</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>346</td>
<td>149</td>
<td>174.0</td>
<td>199</td>
</tr>
<tr>
<td>(output omitted)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>393</td>
<td>180</td>
<td>196.5</td>
<td>213</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>359</td>
<td>144</td>
<td>179.5</td>
<td>215</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>285</td>
<td>125</td>
<td>142.5</td>
<td>160</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>214</td>
<td>102</td>
<td>107.0</td>
<td>112</td>
</tr>
<tr>
<td>(output omitted)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>308</td>
<td>143</td>
<td>154.0</td>
<td>165</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>450</td>
<td>211</td>
<td>225.0</td>
<td>239</td>
</tr>
<tr>
<td>31</td>
<td>62</td>
<td>10351</td>
<td>67</td>
<td>167.0</td>
<td>288</td>
</tr>
</tbody>
</table>
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 = 8720
Number of PSUs = 60  Population size = 98725345
Design df = 29

Linearized
Mean  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
pwgroup: finalwgt
VCE: linearized
Single unit: missing
Strata 1: stratid
SU 1: psuid
FPC 1: <zero>

<table>
<thead>
<tr>
<th>Stratum</th>
<th>#Units included</th>
<th>#Units omitted</th>
<th>#Obs with complete data</th>
<th>#Obs with missing data</th>
<th>#Obs per included Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1*</td>
<td>1</td>
<td>114</td>
<td>266</td>
<td>114</td>
</tr>
<tr>
<td>2</td>
<td>1*</td>
<td>1</td>
<td>98</td>
<td>87</td>
<td>98</td>
</tr>
</tbody>
</table>

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:
. svydes hdresult, finalstage
Survey: Describing final stage sampling units
  pweight: finalwgt
  VCE: linearized
  Single unit: missing
  Strata 1: stratid
  SU 1: psuid
  FPC 1: <zero>

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Unit</th>
<th>#Obs with complete data</th>
<th>#Obs with missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>215</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>114</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>98</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>67</td>
</tr>
</tbody>
</table>

(output omitted)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Unit</th>
<th>#Obs with complete data</th>
<th>#Obs with missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>2</td>
<td>203</td>
<td>8</td>
</tr>
<tr>
<td>31</td>
<td>62</td>
<td>8720</td>
<td>1631</td>
</tr>
</tbody>
</table>

10351

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.

. gen newstr = stratid
. gen 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** — Describe survey data

Survey: Describing final stage sampling units

pweight: finalwgt  
VCE: linearized  
Single unit: missing  
Strata 1: newstr  
SU 1: newpsu  
FPC 1: <zero>

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Unit</th>
<th>#Obs with complete data</th>
<th>#Obs with missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>98</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>215</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>114</td>
<td>51</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>161</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>116</td>
<td>33</td>
</tr>
<tr>
<td>(output omitted)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>180</td>
<td>59</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>203</td>
<td>8</td>
</tr>
<tr>
<td>30</td>
<td>62</td>
<td>8720</td>
<td>1631</td>
</tr>
</tbody>
</table>

10351

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

<table>
<thead>
<tr>
<th>Mean [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linearized</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>hdresult</td>
</tr>
<tr>
<td>49.67141 (.3830147)</td>
</tr>
<tr>
<td>48.88919 (50.45364)</td>
</tr>
</tbody>
</table>

**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/r13/nhanes2b, clear
. svy: mean hdresult
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 31     Number of obs = 8720
Number of PSUs = 60     Population size = 98725345
Design df = 29

Linearized

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hdresult</td>
<td>49.67141</td>
<td>.</td>
</tr>
</tbody>
</table>

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

. svydes 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>

<table>
<thead>
<tr>
<th>Stratum</th>
<th>#Units</th>
<th>#Obs</th>
<th>min</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1*</td>
<td>114</td>
<td>114</td>
<td>114.0</td>
<td>114</td>
</tr>
<tr>
<td>2</td>
<td>1*</td>
<td>98</td>
<td>98</td>
<td>98.0</td>
<td>98</td>
</tr>
</tbody>
</table>

Methods and formulas

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

References


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