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.

Quick start

Describe the stage 1 strata and sampling units
svydescribe

Describe the stage 2 strata and sampling units
svydescribe, stage(2)

Describe the final stage strata and sampling units
svydescribe, finalstage

Describe stage 1 strata, and report on where \( x \) contains missing values
svydescribe \( x \)

Create variable onepsu that identifies strata containing one sampling unit
svydescribe, generate(onepsu)

Show which strata have only one PSU for observations with nonmissing values of \( x \)
svydescribe \( x \), single

Show which strata have only one PSU for observations in the estimation sample
svydescribe if e(sample), single

Menu

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

Syntax

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

Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>stage(#)</td>
<td>sampling stage to describe; default is stage(1)</td>
</tr>
<tr>
<td>finalstage</td>
<td>display information per sampling unit in the final stage</td>
</tr>
<tr>
<td>single</td>
<td>display only the strata with one sampling unit</td>
</tr>
<tr>
<td>generate(newvar)</td>
<td>generate a variable identifying strata with one sampling unit</td>
</tr>
</tbody>
</table>

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

Remarks and examples

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 https://www.stata-press.com/data/r16/nhanes2b
. svyset psuid [pweight=finalwgt], strata(stratid)
        pweight: finalwgt
        VCE: linearized
        Single unit: missing
        Strata 1: stratid
                SU 1: psuid
                FPC 1: <zero>
```
\textbf{svydescribe} will display the strata and PSU arrangement of the dataset.

\begin{verbatim}
. 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

<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>, output omitted</td>
<td>17</td>
<td>2</td>
<td>393</td>
<td>180</td>
<td>196.5</td>
</tr>
<tr>
<td>, output omitted</td>
<td>18</td>
<td>2</td>
<td>359</td>
<td>144</td>
<td>179.5</td>
</tr>
<tr>
<td>, output omitted</td>
<td>20</td>
<td>2</td>
<td>285</td>
<td>125</td>
<td>142.5</td>
</tr>
<tr>
<td>, output omitted</td>
<td>21</td>
<td>2</td>
<td>214</td>
<td>102</td>
<td>107.0</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>10,351</td>
<td>67</td>
<td>167.0</td>
<td>288</td>
</tr>
</tbody>
</table>
\end{verbatim}

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

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

\begin{verbatim}
. svy: mean hdresult
(running \texttt{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

<table>
<thead>
<tr>
<th>Linearized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>hdresult</td>
</tr>
</tbody>
</table>
\end{verbatim}

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
   stratum     #obs     #obs
       1         1*          1           114              266               114.0       114
       2         1*          1           98               87               98.0        98
```

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     #obs with complete data     #obs with missing data
                  #obs     #obs
            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
```

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>

<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>8,720</td>
<td>1,631</td>
</tr>
</tbody>
</table>

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.
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.

```stata
. use https://www.stata-press.com/data/r16/nhanes2b, clear
. svy: mean hresult  
(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

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

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

```stata
. 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>

#Obs per Unit
  
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Units</td>
<td>Obs</td>
<td>min</td>
</tr>
<tr>
<td>1</td>
<td>1*</td>
<td>114</td>
<td>114</td>
</tr>
<tr>
<td>2</td>
<td>1*</td>
<td>98</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