proportion — Estimate proportions

Description

proportion produces estimates of proportions, along with standard errors, for the categories identified by the values in each variable of varlist.

Quick start

Proportions, standard errors, and 95% CIs for each level of v1
proportion v1

Also compute statistics for v2
proportion v1 v2

Treat missing values of v1 as a valid category
proportion v1, missing

As above, for each subpopulation defined by the levels of catvar
proportion v1, missing over(catvar)

Standardizing across strata defined by svar with stratum weight wvar1
proportion v1, stdize(svar) stdweight(wvar1)

Weighting by sampling weight wvar2
proportion v1 [pweight=wvar2]

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Proportions
Syntax

```
proportion varlist [if] [in] [weight] [ , options ]
```

**options**

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>stdize(varname)</td>
<td>variable identifying strata for standardization</td>
</tr>
<tr>
<td>stdweight(varname)</td>
<td>weight variable for standardization</td>
</tr>
<tr>
<td>nostdrescale</td>
<td>do not rescale the standard weight variable</td>
</tr>
<tr>
<td>nolabel</td>
<td>suppress value labels from varlist</td>
</tr>
<tr>
<td>missing</td>
<td>treat missing values like other values</td>
</tr>
<tr>
<td>if/in/over</td>
<td>group over subpopulations defined by varlist; optionally, suppress group labels</td>
</tr>
<tr>
<td>SE/Cluster</td>
<td>vctype may be analytic, _cluster clustvar, bootstrap, or jackknife</td>
</tr>
<tr>
<td>vce(vcetype)</td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>set confidence level; default is level(95)</td>
</tr>
<tr>
<td>level(#)</td>
<td>method to compute limits of confidence intervals; default is citype(logit)</td>
</tr>
<tr>
<td>citype(logit</td>
<td>norm)</td>
</tr>
<tr>
<td>noheader</td>
<td>suppress table header</td>
</tr>
<tr>
<td>nolegend</td>
<td>suppress table legend</td>
</tr>
<tr>
<td>display_options</td>
<td>control column formats and line width</td>
</tr>
<tr>
<td>coeflegend</td>
<td>display legend instead of statistics</td>
</tr>
</tbody>
</table>

Weights are not allowed with the bootstrap prefix; see [R] bootstrap. vce() and weights are not allowed with the svy prefix; see [SVY] svy. fweights, iweights, and pweights are allowed; see [U] 11.1.6 weight. coeflegend does not appear in the dialog box.

See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.

**Options**

**stdize(varname)** specifies that the point estimates be adjusted by direct standardization across the strata identified by varname. This option requires the stdweight() option.

**stdweight(varname)** specifies the weight variable associated with the standard strata identified in the stdize() option. The standardization weights must be constant within the standard strata.

**nostdrescale** prevents the standardization weights from being rescaled within the over() groups. This option requires stdize() but is ignored if the over() option is not specified.

**nolabel** specifies that value labels attached to the variables in varlist be ignored.
missing specifies that missing values in \textit{varlist} be treated as valid categories, rather than omitted from the analysis (the default).

\texttt{if/in/over}  

\texttt{over(varlist [, nolabel])} specifies that estimates be computed for multiple subpopulations, which are identified by the different values of the variables in \textit{varlist}.

When this option is supplied with one variable name, such as \texttt{over(varname)}, the value labels of \textit{varname} are used to identify the subpopulations. If \textit{varname} does not have labeled values (or there are unlabeled values), the values themselves are used, provided that they are nonnegative integers. Noninteger values, negative values, and labels that are not valid Stata names are substituted with a default identifier.

When \texttt{over()} is supplied with multiple variable names, each subpopulation is assigned a unique default identifier.

\texttt{nolabel} requests that value labels attached to the variables identifying the subpopulations be ignored.

\texttt{SE/Cluster}  

\texttt{vce(vcetype)} specifies the type of standard error reported, which includes types that are derived from asymptotic theory (\texttt{analytic}), that allow for intragroup correlation (\texttt{cluster clustvar}), and that use bootstrap or jackknife methods (\texttt{bootstrap, jackknife}); see \texttt{[R] vce_option}.

\texttt{vce(analytic)}, the default, uses the analytically derived variance estimator associated with the sample proportion.

\texttt{level(#)}; see \texttt{[R] estimation options}.

\texttt{citype(logit|normal)} specifies how to compute the limits of confidence intervals.

\texttt{citype(logit)}, the default, uses the logit transformation to compute the limits of confidence intervals.

\texttt{citype(normal)} uses the normal approximation to compute the limits of confidence intervals.

\texttt{noheader} prevents the table header from being displayed. This option implies \texttt{nolegend}.

\texttt{nolegend} prevents the table legend identifying the subpopulations from being displayed.

\texttt{display_options: cformat(\%fmt) and nolstretch}; see \texttt{[R] estimation options}.

The following option is available with \texttt{proportion} but is not shown in the dialog box: \texttt{coeflegend}; see \texttt{[R] estimation options}.
Remarks and examples

Example 1

We can estimate the proportion of each repair rating in auto2.dta:

```
. use http://www.stata-press.com/data/r14/auto2
(1978 Automobile Data)
. proportion rep78
```

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>0.0289855</td>
<td>0.0203446</td>
</tr>
<tr>
<td>Fair</td>
<td>0.115942</td>
<td>0.0388245</td>
</tr>
<tr>
<td>Average</td>
<td>0.4347826</td>
<td>0.0601159</td>
</tr>
<tr>
<td>Good</td>
<td>0.2608696</td>
<td>0.0532498</td>
</tr>
<tr>
<td>Excellent</td>
<td>0.1594203</td>
<td>0.0443922</td>
</tr>
</tbody>
</table>

Here we use the missing option to include missing values as a category of rep78:

```
. proportion rep78, missing
```

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>0.027027</td>
<td>0.0189796</td>
</tr>
<tr>
<td>Fair</td>
<td>0.1081081</td>
<td>0.0363433</td>
</tr>
<tr>
<td>Average</td>
<td>0.4054054</td>
<td>0.0574637</td>
</tr>
<tr>
<td>Good</td>
<td>0.2432432</td>
<td>0.0502154</td>
</tr>
<tr>
<td>Excellent</td>
<td>0.1486486</td>
<td>0.0416364</td>
</tr>
<tr>
<td>_prop_6</td>
<td>0.0675676</td>
<td>0.0293776</td>
</tr>
</tbody>
</table>
Example 2

We can also estimate proportions over groups:

```
. proportion rep78, over(foreign)
```

<table>
<thead>
<tr>
<th>Proportion estimation</th>
<th>Number of obs = 69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor: rep78 = Poor</td>
<td></td>
</tr>
<tr>
<td>Fair: rep78 = Fair</td>
<td></td>
</tr>
<tr>
<td>Average: rep78 = Average</td>
<td></td>
</tr>
<tr>
<td>Good: rep78 = Good</td>
<td></td>
</tr>
<tr>
<td>Excellent: rep78 = Excellent</td>
<td></td>
</tr>
<tr>
<td>Domestic: foreign = Domestic</td>
<td></td>
</tr>
<tr>
<td>Foreign: foreign = Foreign</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Over</th>
<th>Proportion</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domestic</td>
<td>0.0416667</td>
<td>0.0291477</td>
<td>0.0100299 .1572433</td>
</tr>
<tr>
<td>Foreign</td>
<td>0</td>
<td>(no observations)</td>
<td></td>
</tr>
<tr>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domestic</td>
<td>0.1666667</td>
<td>0.0543607</td>
<td>0.0839032 .3039797</td>
</tr>
<tr>
<td>Foreign</td>
<td>0</td>
<td>(no observations)</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domestic</td>
<td>0.5625</td>
<td>0.0723605</td>
<td>0.4169211 .6980553</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.1428571</td>
<td>0.0782461</td>
<td>0.0444941 .3736393</td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domestic</td>
<td>0.1875</td>
<td>0.0569329</td>
<td>0.0986718 .3272601</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.4285714</td>
<td>0.1106567</td>
<td>0.2333786 .6488451</td>
</tr>
<tr>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domestic</td>
<td>0.0416667</td>
<td>0.0291477</td>
<td>0.0100299 .1572433</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.4285714</td>
<td>0.1106567</td>
<td>0.2333786 .6488451</td>
</tr>
</tbody>
</table>
proportion stores the following in e():

Scalars
- e(N): number of observations
- e(N_over): number of subpopulations
- e(N_stdize): number of standard strata
- e(N_clust): number of clusters
- e(k_eq): number of equations in e(b)
- e(df_r): sample degrees of freedom
- e(rank): rank of e(V)

Macros
- e(cmd): proportion
- e(cmdline): command as typed
- e(varlist): varlist
- e(stdize): varname from stdize()
- e(stdweight): varname from stdweight()
- e(vtype): weight type
- e(wexp): weight expression
- e(title): title in estimation output
- e(cluster): name of cluster variable
- e(over): varlist from over()
- e(over_labels): labels from over() variables
- e(over_namelist): names from e(over_labels)
- e(namelist): proportion identifiers
- e(label#): labels from #th variable in varlist
- e(vce): vcetype specified in vce()
- e(vcetype): title used to label Std. Err.
- e(properties): b V
- e(estat_cmd): program used to implement estat
- e(marginsnotok): predictions disallowed by margins

Matrices
- e(b): vector of proportion estimates
- e(V): (co)variance estimates
- e(N): vector of numbers of nonmissing observations
- e(N_stdsum): number of nonmissing observations within the standard strata
- e(p_stdize): standardizing proportions
- e(error): error code corresponding to e(b)

Functions
- e(sample): marks estimation sample

Methods and formulas

Proportions are means of indicator variables; see [R] mean.

Confidence intervals

Confidence intervals for proportions are calculated using a logit transform so that the endpoints lie between 0 and 1. Let \( \hat{p} \) be an estimated proportion and \( \hat{s} \) be an estimate of its standard error. Let

\[
f(\hat{p}) = \ln \left( \frac{\hat{p}}{1-\hat{p}} \right)
\]

be the logit transform of the proportion. In this metric, an estimate of the standard error is

\[
\hat{SE}\{f(\hat{p})\} = f'(\hat{p})\hat{s} = \frac{\hat{s}}{\hat{p}(1-\hat{p})}
\]
Thus a \(100(1 - \alpha)\%\) confidence interval in this metric is

\[
\ln\left( \frac{\hat{p}}{1 - \hat{p}} \right) \pm \frac{t_{1 - \alpha/2, \nu} \hat{s}}{\hat{p}(1 - \hat{p})}
\]

where \(t_{1 - \alpha/2, \nu}\) is the \((1 - \alpha/2)\text{th}\) quantile of Student’s \(t\) distribution with \(\nu\) degrees of freedom. The endpoints of this confidence interval are transformed back to the proportion metric by using the inverse of the logit transform

\[
f^{-1}(y) = \frac{e^y}{1 + e^y}
\]

Hence, the displayed confidence intervals for proportions are

\[
f^{-1}\left\{ \ln\left( \frac{\hat{p}}{1 - \hat{p}} \right) \pm \frac{t_{1 - \alpha/2, \nu} \hat{s}}{\hat{p}(1 - \hat{p})} \right\}
\]

References


Also see

[R] proportion postestimation — Postestimation tools for proportion

[R] mean — Estimate means

[R] ratio — Estimate ratios

[R] total — Estimate totals

[MI] estimation — Estimation commands for use with mi estimate

[SVY] direct standardization — Direct standardization of means, proportions, and ratios

[SVY] poststratification — Poststratification for survey data

[SVY] subpopulation estimation — Subpopulation estimation for survey data

[SVY] svy estimation — Estimation commands for survey data

[SVY] variance estimation — Variance estimation for survey data

[U] 20 Estimation and postestimation commands