Title

stata.com

cmsummarize — Summarize variables by chosen alternatives

Description

`cmsummarize` calculates summary statistics for one or more variables grouped by chosen alternatives.

For panel choice data, `cmsummarize` calculates summary statistics grouped by chosen alternatives and by time.

Quick start

Display the means of `x1` and `x2` grouped by chosen alternatives identified by `depvar` and using `cmset` data

```
cmsummarize x1 x2, choice(depvar)
```

As above, and display sample size, minimum, median, and maximum

```
cmsummarize x1 x2, choice(depvar) statistics(N min median max)
```

For panel choice data, display means of `xvar` grouped by chosen alternatives and time

```
cmsummarize xvar, choice(depvar) time
```

Menu

Statistics > Choice models > Setup and utilities > Summarize variables by chosen alternatives

Syntax

```
cmsummarize varlist [if] [in] [weight] , choice(choicevar) [options]
```

options	Description
Main	
* `choice(`*choicevar*`)`	specify 0/1 variable indicating the chosen alternatives
`statistics(`*statname*`[...])`	report specified statistics
`altwise`	use alternativewise deletion instead of casewise deletion
Reporting	
`format[(%`*fmt*`)]`	display format for statistics; default format is `%9.0g`
`longstub`	put key for statistics (or variable names) on left table stub
`time`	group by time variable (only for panel CM data)
`columns(variables)`	display variables in table columns; the default
`columns(statistics)`	display statistics in table columns

*`choice()` is required.

You must `cmset` your data before using `cmsummarize`; see [CM] **cmset**.

`by` is allowed; see [D] **by**.

`fweights` are allowed; see [U] **11.1.6 weight**.

Options

Main

`choice(`*choicevar*`)` specifies the variable indicating the chosen alternative. *choicevar* must be coded as 0 and 1, with 0 indicating an alternative that was not chosen and 1 indicating the chosen alternative. `choice()` is required.

`statistics(`*statname* `[...])` specifies the statistics to be displayed; the default is equivalent to specifying `statistics(mean)`. (`stats()` is a synonym for `statistics()`.) Multiple statistics may be specified and are separated by white space, such as `statistics(mean sd)`. Available statistics are

statname	Definition	*statname*	Definition
`mean`	mean	`p1`	1st percentile
`count`	count of nonmissing observations	`p5`	5th percentile
`n`	same as `count`	`p10`	10th percentile
`sum`	sum	`p25`	25th percentile
`max`	maximum	`median`	median (same as `p50`)
`min`	minimum	`p50`	50th percentile (same as `median`)
`range`	range = `max` − `min`	`p75`	75th percentile
`sd`	standard deviation	`p90`	90th percentile
`variance`	variance	`p95`	95th percentile
`cv`	coefficient of variation (`sd/mean`)	`p99`	99th percentile
`semean`	standard error of mean ($\texttt{sd}/\sqrt{\texttt{n}}$)	`iqr`	interquartile range = `p75` − `p25`
`skewness`	skewness	`q`	equivalent to specifying `p25 p50 p75`
`kurtosis`	kurtosis		

altwise specifies that alternativewise deletion be used when omitting observations because of missing values in your variables. The default is to use casewise deletion; that is, the entire group of observations making up a case is omitted if any missing values are encountered. This option does not apply to observations that are excluded by the if or in qualifier or the by prefix; these observations are always handled alternativewise regardless of whether altwise is specified.

Reporting

format and format(%*fmt*) specify how the statistics are to be formatted. The default is to use a %9.0g format.

format specifies that each variable's statistics be formatted with the variable's display format; see [D] **format**.

format(%*fmt*) specifies the format to be used for all statistics. The maximum width of the specified format should not exceed nine characters.

longstub specifies that the left stub of the table be made wider so that it can include names of the statistics (or variable names when columns(statistics) is specified) in addition to the categories of the alternatives. The default is to display the names of the statistics (or variable names) in a header.

time groups the statistics by values of the time variable when data are panel choice data. See [CM] **cmset**.

columns(variables | statistics) specifies whether to display variables or statistics in the columns of the table. columns(variables) is the default when more than one variable is specified.

Remarks and examples

stata.com

cmsummarize is a convenience command for displaying summary statistics of one or more variables grouped by chosen alternatives.

The option choice(*choicevar*) is required, where *choicevar* is a 0/1 variable. *choicevar* is typically the dependent variable for choice models with 0/1 dependent variables.

For rank-ordered choice models, such as cmroprobit, using a dependent variable of ranks with choice() will give an error message. To use cmsummarize in this instance, you would have to create a 0/1 variable, such as a variable indicating the highest ranked alternative for each case.

For an overview of other descriptive statistics available for choice model data, see [CM] **Intro 3**.

▷ Example 1: Cross-sectional choice data

Here is an example with cross-sectional choice data. First, we cmset our data:

```
. use https://www.stata-press.com/data/r17/carchoice
(Car choice data)
. cmset consumerid car
note: alternatives are unbalanced across choice sets; choice sets of
      different sizes found.
     Case ID variable: consumerid
Alternatives variable: car
```

These fictitious data represent persons who purchased a car with their choices categorized by the nationality of the manufacturer, American, Japanese, European, or Korean. Statistics are calculated over groups defined by the chosen alternatives, that is, the nationality of car. Second, we type cmsummarize, which by default calculates means. Specifying the variable income, we get the means of income by the nationality of car purchased.

```
. cmsummarize income, choice(purchase)
Statistics by chosen alternatives (purchase = 1)
    income is constant within case
Summary for variables: income
Group variable: _chosen_alternative (purchase = 1)
```

_chosen_alternative	Mean
American	40.52394
Japanese	43.15127
European	45.80462
Korean	35.585
Total	42.05429

The mean income is highest among those that selected European cars.

Third, we specify the option `statistics(N min mean max)` to display the group sample size and the minimum, mean, and maximum of the variables `gender`, `income`, and `dealers`.

```
. cmsummarize gender income dealers, choice(purchase) statistics(N min mean max)
Statistics by chosen alternatives (purchase = 1)
    variables constant within case:
        gender
        income
Summary statistics: N, Min, Mean, Max
Group variable: _chosen_alternative (purchase = 1)
```

_chosen_alternative	gender	income	dealers
American	376	376	376
	0	20.3	2
	.7446809	40.52394	8.143617
	1	69.8	13
Japanese	316	316	316
	0	20.3	1
	.6518987	43.15127	6.25
	1	69.8	12
European	130	130	130
	0	20.3	1
	.8307692	45.80462	3.630769
	1	69.8	7
Korean	40	40	40
	0	20.9	1
	.8	35.585	2.425
	1	69.8	5
Total	862	862	862
	0	20.3	1
	.7262181	42.05429	6.50348
	1	69.8	13

Example 2: Panel choice data

When you have panel choice data, `cmsummarize` is useful to see how summary statistics grouped by chosen alternatives vary by time. Here is an example. First, we `cmset` the data:

```
. use https://www.stata-press.com/data/r17/transport, clear
(Transportation choice data)
. cmset id t alt
note: case identifier _caseid generated from id and t.
note: panel by alternatives identifier _panelaltid generated from id and alt.

                     Panel data: Panels id and time t
               Case ID variable: _caseid
          Alternatives variable: alt
Panel by alternatives variable: _panelaltid (strongly balanced)
                  Time variable: t, 1 to 3
                          Delta: 1 unit

Note: Data have been xtset.
```

Second, we specify the option `time`, which produces statistics grouped by chosen alternatives at each time point. We also specify the formatting for the statistics.

```
. cmsummarize trtime, choice(choice) statistics(median) format(%6.4f) time
Statistics by chosen alternatives (choice = 1)
time t = 1
Summary for variables: trtime
Group variable: _chosen_alternative (choice = 1)
```

_chosen_alternative	p50
Car	0.1764
Public	0.4195
Bicycle	0.5884
Walk	0.8054
Total	0.2316

```
time t = 2
Summary for variables: trtime
Group variable: _chosen_alternative (choice = 1)
```

_chosen_alternative	p50
Car	0.1731
Public	0.3729
Bicycle	0.6562
Walk	0.6671
Total	0.1897

```
time t = 3
Summary for variables: trtime
Group variable: _chosen_alternative (choice = 1)
```

_chosen_alternative	p50
Car	0.1842
Public	0.4345
Bicycle	0.4593
Walk	0.9563
Total	0.2006

If we do not specify the option `time`, statistics are calculated by the groups of chosen alternatives aggregated across time.

```
. cmsummarize trtime, choice(choice) statistics(N min median max) format(%6.0g)
Statistics by chosen alternatives (choice = 1)
Summary for variables: trtime
Group variable: _chosen_alternative (choice = 1)
```

_chosen_alternative	N	Min	p50	Max
Car	981	.1	.1789	.2499
Public	256	.1016	.4171	.7024
Bicycle	145	.102	.573	1.292
Walk	118	.1019	.8126	1.993
Total	1500	.1	.2055	1.993

Also see

[CM] **cmchoiceset** — Tabulate choice sets

[CM] **cmsample** — Display reasons for sample exclusion

[CM] **cmset** — Declare data to be choice model data

[CM] **cmtab** — Tabulate chosen alternatives