Title

stata.com

Intro 3 — Descriptive statistics

Description Remarks and examples Also see

Description

In this entry, we introduce you to four helper commands that let you quickly see some basic attributes of your CM data: `cmchoiceset`, `cmsample`, `cmtab`, and `cmsummarize`.

Remarks and examples

stata.com

Remarks are presented under the following headings:

cmchoiceset: Tabulating choice sets

Let's again use the data that we used in [CM] **Intro 2**.

```
. use https://www.stata-press.com/data/r18/carchoice
(Car choice data)
. list consumerid car purchase gender income if consumerid <= 3,
> sepby(consumerid) abbrev(10)
```

	consumerid	car	purchase	gender	income
1.	1	American	1	Male	46.7
2.	1	Japanese	0	Male	46.7
3.	1	European	0	Male	46.7
4.	1	Korean	0	Male	46.7
5.	2	American	1	Male	26.1
6.	2	Japanese	0	Male	26.1
7.	2	European	0	Male	26.1
8.	2	Korean	0	Male	26.1
9.	3	American	0	Male	32.7
10.	3	Japanese	1	Male	32.7
11.	3	European	0	Male	32.7

The case ID variable is `consumerid`. The alternatives variable is `car`. The 0/1 variable `purchase` indicates the nationality of car purchased. The variables `gender` and `income` are case-specific variables.

We cmset our data:

```
. cmset consumerid car
note: alternatives are unbalanced across choice sets; choice sets of different
      sizes found.
     Case ID variable: consumerid
Alternatives variable: car
```

We use cmchoiceset to see the choice sets:

```
. cmchoiceset
Tabulation of choice-set possibilities
```

Choice set	Freq.	Percent	Cum.
1 2 3	380	42.94	42.94
1 2 3 4	505	57.06	100.00
Total	885	100.00	

```
Note: Total is number of cases.
. label list nation
nation:
           1 American
           2 Japanese
           3 European
           4 Korean
```

There are two choice sets, $\{1,2,3\}$ and $\{1,2,3,4\}$. The value label `nation`, which labels the alternatives variable `car`, shows the correspondence between the numerical values and the nationalities. One choice set includes all four nationalities, and the other includes all nationalities except Korean.

`cmchoiceset` can be used after a `cm` estimation command to see the choice sets in the estimation sample. Here we fit a model using cmclogit and then run `cmchoiceset` restricted to the estimation sample.

```
. cmclogit purchase dealers, casevars(i.gender income)
  (output omitted)
. cmchoiceset if e(sample)
Tabulation of choice-set possibilities
```

Choice set	Freq.	Percent	Cum.
1 2 3	373	43.27	43.27
1 2 3 4	489	56.73	100.00
Total	862	100.00	

```
Note: Total is number of cases.
```

We see that the estimation sample had 862 cases, whereas the earlier `cmchoiceset` output showed that the full sample had 885 cases.

By default, missing values are handled casewise, meaning that any missing value in any observation composing the case causes the entire case to be omitted from the estimation sample. In this example, $885 - 862 = 23$ cases contained missing values.

If you want to omit only observations with missing values and not the entire case, specify the option `altwise`. We refit the model using the `altwise` option and look at the choice sets.

```
. cmclogit purchase dealers, casevars(i.gender income) altwise
  (output omitted)
. cmchoiceset if e(sample)
Tabulation of choice-set possibilities
```

Choice set	Freq.	Percent	Cum.
1 2	2	0.23	0.23
1 2 3	378	42.71	42.94
1 2 3 4	489	55.25	98.19
1 2 4	4	0.45	98.64
1 3	2	0.23	98.87
1 3 4	2	0.23	99.10
2 3	3	0.34	99.44
2 3 4	5	0.56	100.00
Total	885	100.00	

```
Note: Total is number of cases.
```

Handling the missing values alternativewise gives six new choice sets, albeit each with low frequency.

Handling missing values casewise never creates new choice sets. Handling missing values with `altwise` almost always changes the choice sets used in the estimation. You should be aware of the consequences. For instance, a dataset with balanced choice sets will typically become unbalanced when missing values are handled alternativewise. See example 3 in [CM] **cmclogit** for more details.

`cmchoiceset` also creates two-way (and three-way) tabulations. You can tabulate a variable, typically a case-specific one, against choice sets to see whether there is any association between the variable and choice sets. If you have panel data, you can tabulate the choice sets versus time to see whether choice sets change over time. See [CM] **cmchoiceset**.

`cmchoiceset` has a `generate(`*newvar*`)` option, which creates a variable with categories of the choice sets. This variable can be used in the `over()` option of `margins` to compute predicted probabilities and marginal effects separately for each choice set. See example 3 in [CM] **cmchoiceset** for an example.

cmsample: Looking at problem observations

Let's load and try to `cmset` a dataset to which we added some errors.

```
. use https://www.stata-press.com/data/r18/carchoice_errors, clear
(Car choice data with errors)
. cmset consumerid car
at least one choice set has more than one instance of the same alternative
r(459);
```

We get an error and our data are not `cmset`. We need to fix the repeated alternatives in `car`, the alternatives variable. The `cmsample` command can locate these problem observations. But to run `cmsample`, the data must be `cmset`. To do this, we use `cmset` with the `force` option. (Note: `cmsample` is the only command that works after suppressing an error using `cmset, force`. All other `cm` commands will give the same error about repeated alternatives unless the problematic observations are dropped or excluded using an `if` restriction.)

```
. cmset consumerid car, force
note: at least one choice set has more than one instance of the same
      alternative.
     Case ID variable: consumerid
Alternatives variable: car
```

Now we can run `cmsample`. We specify the option `generate(flag)` to create a variable named `flag` that identifies the problem observations.

```
. cmsample, generate(flag)
```

Reason for exclusion	Freq.	Percent	Cum.
observations included	3,153	99.78	99.78
repeated alternatives within case*	7	0.22	100.00
Total	3,160	100.00	

```
* indicates an error
```

`cmsample` produced a table that showed there are seven observations that contain the cases with the repeated alternatives. We can see the problems by listing the observations with `flag != 0`:

```
. list consumerid car flag if flag != 0, sepby(consumerid) abbr(10)
```

	consumerid	car	flag
397.	111	American	repeated alternatives within case*
398.	111	Japanese	repeated alternatives within case*
399.	111	Japanese	repeated alternatives within case*
1035.	290	American	repeated alternatives within case*
1036.	290	Japanese	repeated alternatives within case*
1037.	290	Japanese	repeated alternatives within case*
1038.	290	Korean	repeated alternatives within case*

We will need to fix or drop these cases before we can run other CM commands.

`cmsample` can identify many different problems in your choice data—16 different problems in all! To see its full capabilities, see [CM] **cmsample**.

cmtab: Tabulating chosen alternatives versus other variables

Let's reload our earlier dataset so we are not dealing with a dataset with `cm` errors.

```
. use https://www.stata-press.com/data/r18/carchoice, clear
(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
```

The `cmtab` command requires the `choice(`*varname*`)` option, where *varname* is a 0/1 variable indicating which alternative was chosen. Typically, it is the dependent variable used in a discrete choice model. Typing `cmtab` without any other arguments gives a tabulation of the chosen alternatives:

```
. cmtab, choice(purchase)
Tabulation of chosen alternatives (purchase = 1)
```

Nationality of car	Freq.	Percent	Cum.
American	384	43.39	43.39
Japanese	326	36.84	80.23
European	135	15.25	95.48
Korean	40	4.52	100.00
Total	885	100.00	

Typing `cmtab` with a variable gives a tabulation of that variable versus the chosen alternatives.

```
. cmtab gender, choice(purchase) column
Tabulation for chosen alternatives (purchase = 1)
    gender is constant within case
```

Key
frequency *column percentage*

Nationality of car	Gender: 0 = Female, 1 = Male Female	Male	Total
American	96 40.68	280 44.73	376 43.62
Japanese	110 46.61	206 32.91	316 36.66
European	22 9.32	108 17.25	130 15.08
Korean	8 3.39	32 5.11	40 4.64
Total	236 100.00	626 100.00	862 100.00

We see that in these data, the most popular nationality of car among females was Japanese, with 47% of them purchasing a Japanese car. Among males, American cars were the most popular, with 45% of them buying an American car.

See [CM] **cmtab** for the full capabilities of the command.

cmsummarize: Descriptive statistics for CM variables

The `cmsummarize` command produces descriptive statistics for CM variables. For each variable in the command's *varlist*, it selects observations that correspond to chosen alternatives and displays statistics categorized by the chosen alternatives. The chosen alternatives are specified by the `choice(`*varname*`)` option, which is required, just as it is with `cmtab`.

Here is an example where we display the quartiles of the case-specific variable `income`:

```
. cmsummarize income, choice(purchase) stats(p25 p50 p75) format(%5.1f)
Statistics by chosen alternatives (purchase = 1)
    income is constant within case
Summary for variables: income
Group variable: _chosen_alternative (purchase = 1)
```

_chosen_alternative	p25	p50	p75
American	30.6	42.0	46.6
Japanese	39.0	44.4	48.8
European	40.5	44.6	49.2
Korean	25.4	35.5	44.2
Total	33.0	43.3	46.7

We see that buyers of European cars have the greatest median income and buyers of Korean cars the least compared with buyers of cars of other nationalities.

See [CM] **cmsummarize** for the full capabilities of the command.

Also see

[CM] **Intro 2** — Data layout

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

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

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

[CM] **cmsummarize** — Summarize variables by chosen alternatives

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