# Title

stata.com

**contract —** Make dataset of frequencies and percentages



## Description

`contract` replaces the dataset in memory with a new dataset consisting of all combinations of *varlist* that exist in the data and a new variable that contains the frequency of each combination.

## Quick start

Frequency of each combination of `v1` and `v2` saved in `_freq`

```
contract v1 v2
```

Same as above, but name new frequency variable `newf`

```
contract v1 v2, freq(newf)
```

Add percentage of total in `newp`

```
contract v1 v2, freq(newf) percent(newp)
```

Add cumulative frequency `newcf` and cumulative percentage `newcp`

```
contract v1 v2, freq(newf) percent(newp) cfreq(newcf)  ///
     cpercent(newcp)
```

Frequency of combinations excluding missing values

```
contract v1 v2, nomiss
```

Add combinations with zero observations

```
contract v1 v2, nomiss zero
```

## Menu

Data > Create or change data > Other variable-transformation commands > Make dataset of frequencies

# Syntax

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

| *options* | Description |
|---|---|
| Options | |
| freq(*newvar*) | name of frequency variable; default is _freq |
| cfreq(*newvar*) | create cumulative frequency variable |
| percent(*newvar*) | create percentage variable |
| cpercent(*newvar*) | create cumulative percentage variable |
| float | generate percentage variables as type float |
| format(*format*) | display format for new percentage variables; default is format(%8.2f) |
| zero | include combinations with frequency zero |
| nomiss | drop observations with missing values |

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

# Options

Options

freq(*newvar*) specifies a name for the frequency variable. If not specified, _freq is used.

cfreq(*newvar*) specifies a name for the cumulative frequency variable. If not specified, no cumulative frequency variable is created.

percent(*newvar*) specifies a name for the percentage variable. If not specified, no percentage variable is created.

cpercent(*newvar*) specifies a name for the cumulative percentage variable. If not specified, no cumulative percentage variable is created.

float specifies that the percentage variables specified by percent() and cpercent() will be generated as variables of type float. If float is not specified, these variables will be generated as variables of type double. All generated variables are compressed to the smallest storage type possible without loss of precision; see [D] **compress**.

format(*format*) specifies a display format for the generated percentage variables specified by percent() and cpercent(). If format() is not specified, these variables will have the display format %8.2f.

zero specifies that combinations with frequency zero be included.

nomiss specifies that observations with missing values on any variable in *varlist* be dropped. If nomiss is not specified, all observations possible are used.

# Remarks and examples

stata.com

contract takes the dataset in memory and creates a new dataset containing all combinations of *varlist* that exist in the data and a new variable that contains the frequency of each combination.

Sometimes you may want to collapse a dataset into frequency form. Several observations that have identical values on one or more variables will be replaced by one such observation, together with the frequency of the corresponding set of values. For example, in certain generalized linear models, the frequency of some combination of values is the response variable, so you would need to produce that response variable. The set of covariate values associated with each frequency is sometimes called a covariate class or covariate pattern. Such collapsing is reversible for the variables concerned, because the original dataset can be reconstituted by using `expand` (see [D] **expand**) with the variable containing the frequencies of each covariate class.

## ▷ Example 1

Suppose that we wish to collapse `auto2.dta` to a set of frequencies of the variables `rep78`, which takes values labeled "Poor", "Fair", "Average", "Good", and "Excellent", and `foreign`, which takes values labeled "Domestic" and "Foreign".

```
. use https://www.stata-press.com/data/r18/auto2
(1978 automobile data)
. contract rep78 foreign
. list
```

| | rep78 | foreign | _freq |
|---|---|---|---|
| 1. | Poor | Domestic | 2 |
| 2. | Fair | Domestic | 8 |
| 3. | Average | Domestic | 27 |
| 4. | Average | Foreign | 3 |
| 5. | Good | Domestic | 9 |
| 6. | Good | Foreign | 9 |
| 7. | Excellent | Domestic | 2 |
| 8. | Excellent | Foreign | 9 |
| 9. | . | Domestic | 4 |
| 10. | . | Foreign | 1 |

By default, `contract` uses the variable name `_freq` for the new variable that contains the frequencies. If `_freq` is in use, you will be reminded to specify a new variable name via the `freq()` option.

Specifying the `zero` option requests that combinations with frequency zero also be listed.

```
. use https://www.stata-press.com/data/r18/auto2, clear
(1978 automobile data)
. contract rep78 foreign, zero
. list
```

| | rep78 | foreign | _freq |
|---|---|---|---|
| 1. | Poor | Domestic | 2 |
| 2. | Poor | Foreign | 0 |
| 3. | Fair | Domestic | 8 |
| 4. | Fair | Foreign | 0 |
| 5. | Average | Domestic | 27 |
| 6. | Average | Foreign | 3 |
| 7. | Good | Domestic | 9 |
| 8. | Good | Foreign | 9 |
| 9. | Excellent | Domestic | 2 |
| 10. | Excellent | Foreign | 9 |
| 11. | . | Domestic | 4 |
| 12. | . | Foreign | 1 |

## Acknowledgments

`contract` was written by Nicholas J. Cox (1998) of the Department of Geography at Durham University, UK, who is coeditor of the *Stata Journal* and author of *Speaking Stata Graphics*. The `cfreq()`, `percent()`, `cpercent()`, `float`, and `format()` options were written by Roger Newson of the Imperial College London.

## Reference

Cox, N. J. 1998. dm59: Collapsing datasets to frequencies. *Stata Technical Bulletin* 44: 2–3. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 20–21. College Station, TX: Stata Press.

## Also see

[D] **expand** — Duplicate observations

[D] **collapse** — Make dataset of summary statistics

[D] **duplicates** — Report, tag, or drop duplicate observations