**Description**

`collapse` converts the dataset in memory into a dataset of means, sums, medians, etc. `clist` must refer to numeric variables exclusively.

Note: See [D] `contract` if you want to collapse to a dataset of frequencies.

**Quick start**

Replace dataset in memory with means of `v1` and `v2`

```
collapse v1 v2
```

As above, but calculate statistics separately by each level of `catvar`

```
collapse v1 v2, by(catvar)
```

Dataset of mean, standard deviation, and standard error of the mean of `v1`

```
collapse (mean) mean1=v1 (sd) sd1=v1 (semean) sem1=v1
```

Mean and standard error of the mean for binomial `v2`

```
collapse (mean) mean2=v2 (sebinomial) sem2=v2
```

Frequency, median, and interquartile range of `v1`

```
collapse (count) freq=v1 (p50) p50=v1 (iqr) iqr=v1
```

Weighted and unweighted sum of `v2` using frequency weight `wvar`

```
collapse (sum) weighted=v2 (rawsum) unweighted=v2 [fweight=wvar]
```

**Menu**

Data > Create or change data > Other variable-transformation commands > Make dataset of means, medians, etc.
Syntax

```
collapse clist [if] [in] [weight] [, options]
```

where `clist` is either

- `[(stat)] varlist [ (stat) ] ...`
- `[(stat)] target_var=varname [target_var=varname ...] [ (stat) ] ...`

or any combination of the `varlist` and `target_var` forms, and `stat` is one of

<table>
<thead>
<tr>
<th>Stat</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>means (default)</td>
</tr>
<tr>
<td>median</td>
<td>medians</td>
</tr>
<tr>
<td>p1</td>
<td>1st percentile</td>
</tr>
<tr>
<td>p2</td>
<td>2nd percentile</td>
</tr>
<tr>
<td>...</td>
<td>3rd–49th percentiles</td>
</tr>
<tr>
<td>p50</td>
<td>50th percentile (same as median)</td>
</tr>
<tr>
<td>...</td>
<td>51st–97th percentiles</td>
</tr>
<tr>
<td>p98</td>
<td>98th percentile</td>
</tr>
<tr>
<td>p99</td>
<td>99th percentile</td>
</tr>
<tr>
<td>sd</td>
<td>standard deviations</td>
</tr>
<tr>
<td>semean</td>
<td>standard error of the mean</td>
</tr>
<tr>
<td>sebinomial</td>
<td>standard error of the mean, binomial</td>
</tr>
<tr>
<td>sepoisson</td>
<td>standard error of the mean, Poisson</td>
</tr>
<tr>
<td>sum</td>
<td>sums</td>
</tr>
<tr>
<td>rawsum</td>
<td>sums, ignoring optionally specified weight except observations with a weight of zero are excluded</td>
</tr>
<tr>
<td>count</td>
<td>number of nonmissing observations</td>
</tr>
<tr>
<td>percent</td>
<td>percentage of nonmissing observations</td>
</tr>
<tr>
<td>max</td>
<td>maximums</td>
</tr>
<tr>
<td>min</td>
<td>minimums</td>
</tr>
<tr>
<td>iqr</td>
<td>interquartile range</td>
</tr>
<tr>
<td>first</td>
<td>first value</td>
</tr>
<tr>
<td>last</td>
<td>last value</td>
</tr>
<tr>
<td>firstnm</td>
<td>first nonmissing value</td>
</tr>
<tr>
<td>lastnm</td>
<td>last nonmissing value</td>
</tr>
</tbody>
</table>

If `stat` is not specified, `mean` is assumed.

**options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>by(varlist)</td>
<td>groups over which <code>stat</code> is to be calculated</td>
</tr>
<tr>
<td>cw</td>
<td>casewise deletion instead of all possible observations</td>
</tr>
<tr>
<td>fast</td>
<td>do not restore the original dataset should the user press <code>Break</code>; programmer’s command</td>
</tr>
</tbody>
</table>

`varlist` and `varname` in `clist` may contain time-series operators; see [U] 11.4.4 Time-series varlists.

aweights, fweights, iweights, and pweights are allowed; see [U] 11.1.6 weight, and see Weights below. pweights may not be used with sd, semean, sebinomial, or sepoisson. iweights may not be used with semean, sebinomial, or sepoisson. aweights may not be used with sebinomial or sepoisson.

fast does not appear in the dialog box.

Examples:

- . collapse age educ income, by(state)
- . collapse (mean) age educ (median) income, by(state)
- . collapse (mean) age educ income (median) medinc=income, by(state)
- . collapse (p25) gpa [fw=number], by(year)
**Options**

by(varlist) specifies the groups over which the means, etc., are to be calculated. If this option is not specified, the resulting dataset will contain 1 observation. If it is specified, `varlist` may refer to either string or numeric variables.

cw specifies casewise deletion. If cw is not specified, all possible observations are used for each calculated statistic.

The following option is available with `collapse` but is not shown in the dialog box:

fast specifies that `collapse` not restore the original dataset should the user press `Break`. fast is intended for use by programmers.

**Remarks and examples**

`collapse` takes the dataset in memory and creates a new dataset containing summary statistics of the original data. `collapse` adds meaningful variable labels to the variables in this new dataset. Because the syntax diagram for `collapse` makes using it appear more complicated than it is, `collapse` is best explained with examples.

Remarks are presented under the following headings:

- Introductory examples
- Variablewise or casewise deletion
- Weights
- A final example

**Introductory examples**

➢ **Example 1**

Consider the following artificial data on the grade-point average (gpa) of college students:

```stata
. use https://www.stata-press.com/data/r17/college
. describe
Contains data from https://www.stata-press.com/data/r17/college.dta
Observations: 12
Variables: 4
3 Jan 2020 12:05

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Storage type</th>
<th>Display format</th>
<th>Value label</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpa</td>
<td>float</td>
<td>%9.0g</td>
<td>gpa for this year</td>
</tr>
<tr>
<td>hour</td>
<td>int</td>
<td>%9.0g</td>
<td>Total academic hours</td>
</tr>
<tr>
<td>year</td>
<td>int</td>
<td>%9.0g</td>
<td>1 = freshman, 2 = sophomore, 3 = junior, 4 = senior</td>
</tr>
<tr>
<td>number</td>
<td>int</td>
<td>%9.0g</td>
<td>number of students</td>
</tr>
</tbody>
</table>
```

Sorted by: year
To obtain a dataset containing the 25th percentile of gpa’s for each year, we type

\[
.\text{collapse (p25) gpa [fw=number], by(year)}
\]

We used frequency weights.

Next we want to create a dataset containing the mean of gpa and hour for each year. We do not have to type (mean) to specify that we want the mean because the mean is reported by default.

\[
.\text{use https://www.stata-press.com/data/r17/college, clear}
.\text{collapse gpa hour [fw=number], by(year)}
.\text{list}
\]

Now we want to create a dataset containing the mean and median of gpa and hour, and we want the median of gpa and hour to be stored as variables medgpa and medhour, respectively.

\[
.\text{use https://www.stata-press.com/data/r17/college, clear}
.\text{collapse (mean) gpa hour (median) medgpa=gpa medhour=hour [fw=num], by(year)}
.\text{list}
\]

Here we want to create a dataset containing a count of gpa and hour and the minimums of gpa and hour. The minimums of gpa and hour will be stored as variables mingpa and minhour, respectively.
. use https://www.stata-press.com/data/r17/college, clear
. collapse (count) gpa hour (min) mingpa=gpa minhour=hour [fw=num], by(year)
. list

<table>
<thead>
<tr>
<th>year</th>
<th>gpa</th>
<th>hour</th>
<th>mingpa</th>
<th>minhour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>18</td>
<td>2.1</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>12</td>
<td>2.5</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>9</td>
<td>2.2</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>7</td>
<td>2.9</td>
<td>31</td>
</tr>
</tbody>
</table>

Now we replace the values of gpa in 3 of the observations with missing values.

. use https://www.stata-press.com/data/r17/college, clear
. replace gpa = . in 2/4
   (3 real changes made, 3 to missing)
. list, sep(4)

<table>
<thead>
<tr>
<th>gpa</th>
<th>hour</th>
<th>year</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>30</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>.</td>
<td>34</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>.</td>
<td>28</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>.</td>
<td>30</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3.8</td>
<td>29</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2.5</td>
<td>30</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2.9</td>
<td>35</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3.7</td>
<td>30</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>35</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3.3</td>
<td>33</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3.4</td>
<td>32</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2.9</td>
<td>31</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

If we now want to list the data containing the mean of gpa and hour for each year, `collapse` uses all observations on hour for year = 1, even though gpa is missing for observations 1–3.

. collapse gpa hour [fw=num], by(year)
. list

<table>
<thead>
<tr>
<th>year</th>
<th>gpa</th>
<th>hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.2</td>
<td>29.44444</td>
</tr>
<tr>
<td>2</td>
<td>2.991667</td>
<td>31.83333</td>
</tr>
<tr>
<td>3</td>
<td>3.233333</td>
<td>32.11111</td>
</tr>
<tr>
<td>4</td>
<td>3.257143</td>
<td>31.71428</td>
</tr>
</tbody>
</table>
If we repeat this process but specify the `cw` option, `collapse` ignores all observations that have missing values.

```
. use https://www.stata-press.com/data/r17/college, clear
. replace gpa = . in 2/4
   (3 real changes made, 3 to missing)
. collapse (mean) gpa hour [fw=num], by(year) cw
. list
```

<table>
<thead>
<tr>
<th>year</th>
<th>gpa</th>
<th>hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.2</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>2.991667</td>
<td>31.83333</td>
</tr>
<tr>
<td>3</td>
<td>3.233333</td>
<td>32.11111</td>
</tr>
<tr>
<td>4</td>
<td>3.257143</td>
<td>31.71428</td>
</tr>
</tbody>
</table>

Example 2

We have individual-level data from a census in which each observation is a person. Among other variables, the dataset contains the numeric variables `age`, `educ`, and `income` and the string variable `state`. We want to create a 50-observation dataset containing the means of `age`, education, and income for each state.

```
. collapse age educ income, by(state)
```

The resulting dataset contains means because `collapse` assumes that we want means if we do not specify otherwise. To make this explicit, we could have typed

```
. collapse (mean) age educ income, by(state)
```

Had we wanted the mean for `age` and `educ` and the median for `income`, we could have typed

```
. collapse (mean) age educ (median) income, by(state)
```

or if we had wanted the mean for `age` and `educ` and both the mean and the median for `income`, we could have typed

```
. collapse (mean) age educ income (median) medinc=income, by(state)
```

This last dataset will contain three variables containing means—`age`, `educ`, and `income”—and one variable containing the median of income—`medinc`. Because we typed `(median) medinc=income`, Stata knew to find the median for income and to store those in a variable named `medinc`. This renaming convention is necessary in this example because a variable named `income` containing the mean is also being created.
Variablewise or casewise deletion

Example 3

Let’s assume that in our census data, we have 25,000 persons for whom age is recorded but only 15,000 for whom income is recorded; that is, income is missing for 10,000 observations. If we want summary statistics for age and income, collapse will, by default, use all 25,000 observations when calculating the summary statistics for age. If we prefer that collapse use only the 15,000 observations for which income is not missing, we can specify the cw (casewise) option:

```
    . collapse (mean) age income (median) medinc=income, by(state) cw
```

Weights

collapse allows all four weight types; the default is aweights. Weight normalization affects only the `sum`, `count`, `sd`, `semean`, and `sebinomial` statistics.

Let $j$ index observations and $i$ index by-groups. Here are the definitions for `count` and `sum` with weights:

**count:**
- unweighted: $N_i$, the number of observations in group $i$
- aweight: $N_i$, the number of observations in group $i$
- fweight, iweight, pweight: $\sum w_j$, the sum of the weights over observations in group $i$

**sum:**
- unweighted: $\sum x_j$, the sum of $x_j$ over observations in group $i$
- aweight: $\sum v_j x_j$ over observations in group $i$; $v_j =$ weights normalized to sum to $N_i$
- fweight, iweight, pweight: $\sum w_j x_j$ over observations in group $i$

When the by() option is not specified, the entire dataset is treated as one group.

The `sd` statistic with weights returns the square root of the bias-corrected variance, which is based on the factor $\sqrt{N_i/(N_i-1)}$, where $N_i$ is the number of observations. Statistics `sd`, `semean`, `sebinomial`, and `sepoisson` are not allowed with weighted data. Otherwise, the statistic is changed by the weights through the computation of the weighted count, as outlined above.

For instance, consider a case in which there are 25 observations in the dataset and a weighting variable that sums to 57. In the unweighted case, the weight is not specified, and the count is 25. In the analytically weighted case, the count is still 25; the scale of the weight is irrelevant. In the frequency-weighted case, however, the count is 57, the sum of the weights.

The `rawsum` statistic with `aweights` ignores the weight, with one exception: observations with zero weight will not be included in the sum.
Example 4

Using our same census data, suppose that instead of starting with individual-level data and aggregating to the state level, we started with state-level data and wanted to aggregate to the region level. Also assume that our dataset contains pop, the population of each state.

To obtain unweighted means and medians of age and income, by region, along with the total population, we could type

```
collapse (mean) age income (median) medage=age medinc=income (sum) pop,
> by(region)
```

To obtain weighted means and medians of age and income, by region, along with the total population and using frequency weights, we could type

```
collapse (mean) age income (median) medage=age medinc=income (count) pop
> [fweight=pop], by(region)
```

Note: Specifying (sum) pop would not have worked because that would have yielded the pop-weighted sum of pop. Specifying (count) age would have worked as well as (count) pop because count merely counts the number of nonmissing observations. The counts here, however, are frequency-weighted and equal the sum of pop.

To obtain the same mean and medians as above, but using analytic weights, we could type

```
collapse (mean) age income (median) medage=age medinc=income (rawsum) pop
> [aweight=pop], by(region)
```

Note: Specifying (count) pop would not have worked because, with analytic weights, count would count numbers of physical observations. Specifying (sum) pop would not have worked because sum would calculate weighted sums (with a normalized weight). The rawsum function, however, ignores the weights and sums only the specified variable, with one exception: observations with zero weight will not be included in the sum. rawsum would have worked as the solution to all three cases.

A final example

Example 5

We have census data containing information on each state’s median age, marriage rate, and divorce rate. We want to form a new dataset containing various summary statistics, by region, of the variables:
use https://www.stata-press.com/data/r17/census5, clear
(1980 Census data by state)

describe
Contains data from https://www.stata-press.com/data/r17/census5.dta
Observations: 50 1980 Census data by state
Variables: 7 6 Apr 2020 15:43

Variable Storage Display Value
name type format label Variable label

state str14 %14s State
state2 str2 %-2s Two-letter state abbreviation
region int %8.0g cenreg Census region
pop long %10.0g Population
median_age float %9.2f Median age
marriage_rate float %9.0g
divorce_rate float %9.0g

Sorted by: region

collapse (median) median_age marriage divorce (mean) avgmrate=marriage > avgdrate=divorce [aw=pop], by(region)

list

<table>
<thead>
<tr>
<th>region</th>
<th>median age</th>
<th>marriage rate</th>
<th>divorce rate</th>
<th>avgm rate</th>
<th>avgd rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>31.90</td>
<td>0.0080657</td>
<td>0.0035295</td>
<td>0.0081472</td>
<td>0.0035359</td>
</tr>
<tr>
<td>N Ctrl</td>
<td>29.90</td>
<td>0.0093821</td>
<td>0.0048636</td>
<td>0.0096701</td>
<td>0.004961</td>
</tr>
<tr>
<td>South</td>
<td>29.60</td>
<td>0.0112609</td>
<td>0.0065792</td>
<td>0.0117082</td>
<td>0.0059439</td>
</tr>
<tr>
<td>West</td>
<td>29.90</td>
<td>0.0089093</td>
<td>0.0056423</td>
<td>0.0125199</td>
<td>0.0063464</td>
</tr>
</tbody>
</table>

describe
Contains data
Observations: 4 1980 Census data by state
Variables: 6

Sorted by: region

Note: Dataset has changed since last saved.

Acknowledgment

We thank David Roodman of the Open Philanthropy Project for writing collapse2, which inspired several features in collapse.
Also see

[D] contract — Make dataset of frequencies and percentages

[D] egen — Extensions to generate

[D] statsby — Collect statistics for a command across a by list

[R] summarize — Summary statistics