**summarize — Summary statistics**

<table>
<thead>
<tr>
<th>Description</th>
<th>Quick start</th>
<th>Menu</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>Remarks and examples</td>
<td>Stored results</td>
<td>Methods and formulas</td>
</tr>
<tr>
<td>References</td>
<td>Also see</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Description**

`summarize` calculates and displays a variety of univariate summary statistics. If no `varlist` is specified, summary statistics are calculated for all the variables in the dataset.

**Quick start**

Basic summary statistics for continuous variable `v1`

```
summarize v1
```

Same as above, and include `v2` and `v3`

```
summarize v1-v3
```

Same as above, and provide additional detail about the distribution

```
summarize v1-v3, detail
```

Summary statistics reported separately for each level of `catvar`

```
by catvar: summarize v1
```

With frequency weight `wvar`

```
summarize v1 [fweight=wvar]
```

**Menu**

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

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

**options**

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main</strong></td>
</tr>
<tr>
<td><strong>detail</strong></td>
</tr>
<tr>
<td><strong>meanonly</strong></td>
</tr>
<tr>
<td><strong>format</strong></td>
</tr>
<tr>
<td><strong>separator(#)</strong></td>
</tr>
<tr>
<td><strong>display_options</strong></td>
</tr>
</tbody>
</table>

*varlist* may contain factor variables; see [U] 11.4.3 Factor variables.
*varlist* may contain time-series operators; see [U] 11.4.4 Time-series varlists.
by, rolling, and statsby are allowed; see [U] 11.1.10 Prefix commands.
aweights, fweights, and iweights are allowed. However, iweights may not be used with the detail option; see [U] 11.1.6 weight.

**Options**

- **Main**
  - detail produces additional statistics, including skewness, kurtosis, the four smallest and four largest values, and various percentiles.
  - meanonly, which is allowed only when detail is not specified, suppresses the display of results and calculation of the variance. Ado-file writers will find this useful for fast calls.
  - format requests that the summary statistics be displayed using the display formats associated with the variables rather than the default g display format; see [U] 12.5 Formats: Controlling how data are displayed.
  - separator(#) specifies how often to insert separation lines into the output. The default is separator(5), meaning that a line is drawn after every five variables. separator(10) would draw a line after every 10 variables. separator(0) suppresses the separation line.
- **display_options**: vsquish, noemptycells, baselevels, allbaselevels, nofylabel, fvwrap(#), and fvwrapon(style); see [R] Estimation options.

**Remarks and examples**

`summarize` can produce two different sets of summary statistics. Without the detail option, the number of nonmissing observations, the mean and standard deviation, and the minimum and maximum values are presented. With detail, the same information is presented along with the variance, skewness, and kurtosis; the four smallest and four largest values; and the 1st, 5th, 10th, 25th, 50th (median), 75th, 90th, 95th, and 99th percentiles.

Also see [R] ci for calculating the standard error and confidence intervals of the mean.
Example 1: summarize with the separator() option

We have data containing information on various automobiles, among which is the variable mpg, the mileage rating. We can obtain a quick summary of the mpg variable by typing

```
   . use https://www.stata-press.com/data/r16/auto2
   (1978 Automobile Data)
   . summarize mpg
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>74</td>
<td>21.2973</td>
<td>5.785503</td>
<td>12</td>
<td>41</td>
</tr>
</tbody>
</table>

We see that we have 74 observations. The mean of mpg is 21.3 miles per gallon, and the standard deviation is 5.79. The minimum is 12, and the maximum is 41.

If we had not specified the variable (or variables) we wanted to summarize, we would have obtained summary statistics on all the variables in the dataset:

```
   . summarize, separator(4)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>make</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price</td>
<td>74</td>
<td>6165.26</td>
<td>2949.496</td>
<td>3291</td>
<td>15906</td>
</tr>
<tr>
<td>mpg</td>
<td>74</td>
<td>21.2973</td>
<td>5.785503</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>rep78</td>
<td>69</td>
<td>3.40579</td>
<td>0.9899323</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>headroom</td>
<td>74</td>
<td>2.993243</td>
<td>.8459948</td>
<td>1.5</td>
<td>5</td>
</tr>
<tr>
<td>trunk</td>
<td>74</td>
<td>13.75676</td>
<td>4.277404</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>weight</td>
<td>74</td>
<td>3019.469</td>
<td>777.1936</td>
<td>1760</td>
<td>4840</td>
</tr>
<tr>
<td>length</td>
<td>74</td>
<td>187.9324</td>
<td>22.26634</td>
<td>142</td>
<td>233</td>
</tr>
<tr>
<td>turn</td>
<td>74</td>
<td>39.64865</td>
<td>4.399354</td>
<td>31</td>
<td>51</td>
</tr>
<tr>
<td>displacement</td>
<td>74</td>
<td>197.2973</td>
<td>91.83722</td>
<td>79</td>
<td>425</td>
</tr>
<tr>
<td>gear_ratio</td>
<td>74</td>
<td>3.014865</td>
<td>.4562871</td>
<td>2.19</td>
<td>3.89</td>
</tr>
<tr>
<td>foreign</td>
<td>74</td>
<td>.2972973</td>
<td>.4601885</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

There are only 69 observations on rep78, so some of the observations are missing. There are no observations on make because it is a string variable.
The idea of the mean is quite old (Plackett 1958), but its extension to a scheme of moment-based measures was not done until the end of the 19th century. Between 1893 and 1905, Pearson discussed and named the standard deviation, skewness, and kurtosis, but he was not the first to use any of these. Thiele (1889), in contrast, had earlier firmly grasped the notion that the \( m_r \) provide a systematic basis for discussing distributions. However, even earlier anticipations can also be found. For example, Euler in 1778 used \( m_2 \) and \( m_3 \) in passing in a treatment of estimation (Hald 1998, 87), but seemingly did not build on that.

Similarly, the idea of the median is quite old. The history of the interquartile range is tangled up with that of the probable error, a long-popular measure. Extending this in various ways to a more general approach based on quantiles (to use a later term) occurred to several people in the nineteenth century. Galton (1875) is a nice example, particularly because he seems so close to the key idea of the quantiles as a function, which took another century to reemerge strongly.

Thorvald Nicolai Thiele (1838–1910) was a Danish scientist who worked in astronomy, mathematics, actuarial science, and statistics. He made many pioneering contributions to statistics, several of which were overlooked until recently. Thiele advocated graphical analysis of residuals checking for trends, symmetry of distributions, and changes of sign, and he even warned against overinterpreting such graphs.

> Example 2: summarize with the detail option

The detail option provides all the information of a normal `summarize` and more. The format of the output also differs, as shown here:

```
. summarize mpg, detail

                  Mileage (mpg)
                      Percentiles       Smallest
                      1%       12     12
                      5%       14     12
                     10%      14     14       Obs       74
                     25%      18     14       Sum of Wgt.     74
                     50%      20                Mean        21.2973
                     75%      25                Std. Dev.    5.785503
                    90%      29                Variance    33.47205
                    95%      34                Skewness     0.9487176
                    99%      41                Kurtosis    3.975005

```

As in the previous example, we see that the mean of `mpg` is 21.3 miles per gallon and that the standard deviation is 5.79. We also see the various percentiles. The median of `mpg` (the 50th percentile) is 20 miles per gallon. The 25th percentile is 18, and the 75th percentile is 25.

When we performed `summarize`, we learned that the minimum and maximum were 12 and 41, respectively. We now see that the four smallest values in our dataset are 12, 12, 14, and 14. The four largest values are 34, 35, 35, and 41. The skewness of the distribution is 0.95, and the kurtosis is 3.98. (A normal distribution would have a skewness of 0 and a kurtosis of 3.)

Skewness is a measure of the lack of symmetry of a distribution. If the distribution is symmetric, the coefficient of skewness is 0. If the coefficient is negative, the median is usually greater than the mean and the distribution is said to be skewed left. If the coefficient is positive, the median is usually less than the mean and the distribution is said to be skewed right. Kurtosis (from the Greek kurtosis, meaning curvature) is a measure of peakedness of a distribution. The smaller the coefficient of kurtosis, the flatter the distribution. The normal distribution has a coefficient of kurtosis of 3 and provides a convenient benchmark.
Technical note

The convention of calculating the median of an even number of values by averaging the central two order statistics is of long standing. (That is, given 8 values, average the 4th and 5th smallest values, or given 42, average the 21st and 22nd smallest.) Stigler (1977) filled a much-needed gap in the literature by naming such paired central order statistics as “comedians”, although it remains unclear how far he was joking.

Example 3: summarize with the by prefix

`summarize` can usefully be combined with the `by varlist:` prefix. In our dataset, we have a variable, `foreign`, that distinguishes foreign and domestic cars. We can obtain summaries of `mpg` and `weight` within each subgroup by typing

```
    . by foreign: summarize mpg weight
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>52</td>
<td>19.82692</td>
<td>4.743297</td>
<td>12</td>
<td>34</td>
</tr>
<tr>
<td>weight</td>
<td>52</td>
<td>3317.115</td>
<td>695.3637</td>
<td>1800</td>
<td>4840</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>22</td>
<td>24.77273</td>
<td>6.611187</td>
<td>14</td>
<td>41</td>
</tr>
<tr>
<td>weight</td>
<td>22</td>
<td>2315.909</td>
<td>433.0035</td>
<td>1760</td>
<td>3420</td>
</tr>
</tbody>
</table>

Domestic cars in our dataset average 19.8 miles per gallon, whereas foreign cars average 24.8.

Because `by varlist:` can be combined with `summarize`, it can also be combined with `summarize, detail:`
. by foreign: summarize mpg, detail

-> foreign = Domestic

Mileage (mpg)

Percentiles Smallest
1% 12 12
5% 14 12
10% 14 14 Obs 52
25% 16.5 14 Sum of Wgt. 52
50% 19 Mean 19.82692
Largest Std. Dev. 4.743297
75% 22 28
90% 26 29 Variance 22.49887
95% 29 30 Skewness .7712432
99% 34 34 Kurtosis 3.441459

-> foreign = Foreign

Mileage (mpg)

Percentiles Smallest
1% 14 14
5% 17 17
10% 17 17 Obs 22
25% 21 18 Sum of Wgt. 22
50% 24.5 Mean 24.77273
Largest Std. Dev. 6.611187
75% 28 31
90% 35 35 Variance 43.70779
95% 35 35 Skewness .657329
99% 41 41 Kurtosis 3.10734

Technical note

summarize respects display formats if we specify the format option. When we type summarize price weight, we obtain

. summarize price weight

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>74</td>
<td>6,165.26</td>
<td>2,949.496</td>
<td>3291</td>
<td>15,906</td>
</tr>
<tr>
<td>weight</td>
<td>74</td>
<td>3,019.46</td>
<td>777.1936</td>
<td>1,760</td>
<td>4,840</td>
</tr>
</tbody>
</table>

The display is accurate but is not as aesthetically pleasing as we may wish, particularly if we plan to use the output directly in published work. By placing formats on the variables, we can control how the table appears:

. format price weight %9.2fc

. summarize price weight, format

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>74</td>
<td>6,165.26</td>
<td>2,949.50</td>
<td>3,291</td>
<td>15,906</td>
</tr>
<tr>
<td>weight</td>
<td>74</td>
<td>3,019.46</td>
<td>777.19</td>
<td>1,760</td>
<td>4,840</td>
</tr>
</tbody>
</table>
If you specify a weight (see [U] 11.1.6 weight), each observation is multiplied by the value of the weighting expression before the summary statistics are calculated so that the weighting expression is interpreted as the discrete density of each observation.

Example 4: summarize with factor variables

You can also use `summarize` to obtain summary statistics for factor variables. For example, if you type

```
. summarize i.rep78
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>69</td>
<td>.0289855</td>
<td>.1689948</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Fair</td>
<td>69</td>
<td>.115942</td>
<td>.3225009</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>69</td>
<td>.4347826</td>
<td>.4993602</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Good</td>
<td>69</td>
<td>.2608696</td>
<td>.4423259</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Excellent</td>
<td>69</td>
<td>.1594203</td>
<td>.3687494</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

you obtain the sample proportions for the five levels of the `rep78` variable. For example, 11.6% of the 69 cars with nonmissing values of `rep78` have a fair repair record.

We could have used `tabulate oneway rep78` to obtain the sample proportions along with the cumulative proportions. Alternatively, we could have used `proportions rep78` to obtain the sample proportions along with the standard errors of the proportions instead of the standard deviations of the proportions.

Example 5: summarize with weights

We have 1980 census data on each of the 50 states. Included in our variables is `medage`, the median age of the population of each state. If we type `summarize medage`, we obtain unweighted statistics:

```
. use https://www.stata-press.com/data/r16/census, clear
(1980 Census data by state)
. summarize medage
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>medage</td>
<td>50</td>
<td>29.54</td>
<td>1.693445</td>
<td>24.2</td>
<td>34.7</td>
</tr>
</tbody>
</table>

Also among our variables is `pop`, the population in each state. Typing `summarize medage [w=pop]` produces population-weighted statistics:
The number listed under Weight is the sum of the weighting variable, pop, indicating that there are roughly 226 million people in the United States. The pop-weighted mean of medage is 30.11 (compared with 29.54 for the unweighted statistic), and the weighted standard deviation is 1.67 (compared with 1.69).

Example 6: summarize with weights and the detail option

We can obtain detailed summaries of weighted data as well. When we do this, all the statistics are weighted, including the percentiles.

```
. summarize medage [w=pop], detail
(analytic weights assumed)
```

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>27.1</td>
</tr>
<tr>
<td>5%</td>
<td>27.7</td>
</tr>
<tr>
<td>10%</td>
<td>28.2</td>
</tr>
<tr>
<td>25%</td>
<td>29.2</td>
</tr>
<tr>
<td>50%</td>
<td>29.9</td>
</tr>
<tr>
<td>75%</td>
<td>30.9</td>
</tr>
<tr>
<td>90%</td>
<td>32.1</td>
</tr>
<tr>
<td>95%</td>
<td>32.2</td>
</tr>
<tr>
<td>99%</td>
<td>34.7</td>
</tr>
</tbody>
</table>

```
Median age

Percentiles Smallest
1% 27.1 24.2
5% 27.7 26.1
10% 28.2 27.1  Obs 50
25% 29.2 27.4 Sum of Wgt. 225907472
50% 29.9 Mean 30.11047
75% 30.9 Largest Std. Dev. 1.66933
90% 32.1 Variance 2.786661
95% 32.2 Skewness .5281972
99% 34.7 Kurtosis 4.494223
```

Technical note

If you are writing a program and need to access the mean of a variable, the meanonly option provides for fast calls. For example, suppose that your program reads as follows:

```stata
program mean
    summarize '1', meanonly
    display " mean = " r(mean)
end
```

The result of executing this is

```
. use https://www.stata-press.com/data/r16/auto2
(1978 Automobile Data)
. mean price
    mean = 6165.2568
```

Summary statistics

Descriptive statistics in Stata

Stored results

`summarize` stores the following in `r()`:

Scalars

- `r(N)` number of observations
- `r(mean)` mean
- `r(skewness)` skewness (detail only)
- `r(max)` maximum
- `r(sum_w)` sum of the weights
- `r(p1)` 1st percentile (detail only)
- `r(p5)` 5th percentile (detail only)
- `r(p10)` 10th percentile (detail only)
- `r(p25)` 25th percentile (detail only)
- `r(p50)` 50th percentile (detail only)
- `r(p75)` 75th percentile (detail only)
- `r(p90)` 90th percentile (detail only)
- `r(p95)` 95th percentile (detail only)
- `r(p99)` 99th percentile (detail only)
- `r(kurtosis)` kurtosis (detail only)
- `r(var)` variance
- `r(sum)` sum of variable
- `r(sd)` standard deviation

Methods and formulas

Let \( x \) denote the variable on which we want to calculate summary statistics, and let \( x_i, i = 1, \ldots, n, \) denote an individual observation on \( x \). Let \( v_i \) be the weight, and if no weight is specified, define \( v_i = 1 \) for all \( i \).

Define \( V \) as the sum of the weight:

\[
V = \sum_{i=1}^{n} v_i
\]

Define \( w_i \) to be \( v_i \) normalized to sum to \( n \), \( w_i = v_i(n/V) \).

The mean, \( \bar{x} \), is defined as

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} w_i x_i
\]

The variance, \( s^2 \), is defined as

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} w_i (x_i - \bar{x})^2
\]

The standard deviation, \( s \), is defined as \( \sqrt{s^2} \).

Define \( m_r \) as the \( r \)th moment about the mean \( \bar{x} \):

\[
m_r = \frac{1}{n} \sum_{i=1}^{n} w_i (x_i - \bar{x})^r
\]

The coefficient of skewness is then defined as \( m_3 m_2^{-3/2} \). The coefficient of kurtosis is defined as \( m_4 m_2^{-2} \).

Let \( x(i) \) refer to the \( x \) in ascending order, and let \( w(i) \) refer to the corresponding weights of \( x(i) \). The four smallest values are \( x(1), x(2), x(3), \) and \( x(4) \). The four largest values are \( x(n), x(n-1), x(n-2), \) and \( x(n-3) \).
To obtain the \( p \)th percentile, which we will denote as \( x[p] \), let \( P = np/100 \). Let

\[
W(i) = \sum_{j=1}^{i} w(j)
\]

Find the first index \( i \) such that \( W(i) > P \). The \( p \)th percentile is then

\[
x[p] = \begin{cases} 
\frac{x(i-1) + x(i)}{2} & \text{if } W(i-1) = P \\
x(i) & \text{otherwise}
\end{cases}
\]

References


Also see

[R] ameans — Arithmetic, geometric, and harmonic means
[R] centile — Report centile and confidence interval
[R] ci — Confidence intervals for means, proportions, and variances
[R] mean — Estimate means
[R] proportion — Estimate proportions
[R] ratio — Estimate ratios
[R] table — Flexible table of summary statistics
[R] tabstat — Compact table of summary statistics
[R] tabulate, summarize() — One- and two-way tables of summary statistics
[R] total — Estimate totals
[D] codebook — Describe data contents
[D] describe — Describe data in memory or in file
[D] inspect — Display simple summary of data’s attributes
[ST] stsum — Summarize survival-time data
[SVY] svy estimation — Estimation commands for survey data
[XT] xtsum — Summarize xt data