Title

stata.com

summarize — Summary statistics

Description | Quick start | Menu | Syntax
Options | Remarks and examples | Stored results | Methods and formulas
References | Also see

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	Description
Main	
`detail`	display additional statistics
`meanonly`	suppress the display; calculate only the mean; programmer's option
`format`	use variable's display format
`separator(#)`	draw separator line after every # variables; default is `separator(5)`
display_options	control spacing, line width, and base and empty cells

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`, `collect`, `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`, `nofvlabel`, `fvwrap(#)`, and `fvwrapon(`*style*`)`; see [R] **Estimation options**.

Remarks and examples

stata.com

`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/r17/auto2
(1978 automobile data)
. summarize mpg
```

Variable	Obs	Mean	Std. dev.	Min	Max
mpg	74	21.2973	5.785503	12	41

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)
```

Variable	Obs	Mean	Std. dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

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
                        Largest       Std. dev.       5.785503
75%           25             34
90%           29             35       Variance       33.47205
95%           34             35       Skewness       .9487176
99%           41             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 *kyrtosis*, 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
```

```
-> foreign = Domestic
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         mpg |         52    19.82692    4.743297        12         34
      weight |         52    3317.115    695.3637      1800       4840
-------------------------------------------------------------------------------
-> foreign = Foreign
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         mpg |         22    24.77273    6.611187        14         41
      weight |         22    2315.909    433.0035      1760       3420
```

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
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906
      weight |         74    3019.459    777.1936       1760       4840
```

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
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6,165.26    2,949.50   3,291.00  15,906.00
      weight |         74    3,019.46      777.19   1,760.00   4,840.00
```

❑

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
```

Variable	Obs	Mean	Std. dev.	Min	Max
rep78					
Poor	69	.0289855	.1689948	0	1
Fair	69	.115942	.3225009	0	1
Average	69	.4347826	.4993602	0	1
Good	69	.2608696	.4423259	0	1
Excellent	69	.1594203	.3687494	0	1

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/r17/census, clear
(1980 Census data by state)
. summarize medage
```

Variable	Obs	Mean	Std. dev.	Min	Max
medage	50	29.54	1.693445	24.2	34.7

Also among our variables is `pop`, the population in each state. Typing `summarize medage [w=pop]` produces population-weighted statistics:

```
. summarize medage [w=pop]
(analytic weights assumed)
    Variable |       Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
      medage |        50   225907472    30.11047    1.66933        24.2       34.7
```

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)
                          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
                       Largest         Std. dev.      1.66933
75%         30.9              32
90%         32.1            32.1       Variance       2.786661
95%         32.2            32.2       Skewness       .5281972
99%         34.7            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:

```
program mean
        summarize ‘1’, meanonly
        display "  mean = " r(mean)
end
```

The result of executing this is

```
. use https://www.stata-press.com/data/r17/auto2
(1978 automobile data)
. mean price
  mean = 6165.2568
```

❑

Video example

Descriptive statistics in Stata

Stored results

summarize stores the following in r():

Scalars

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

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*, $\overline{x}$, is defined as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} w_i x_i$$

The *variance*, s^2, is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} w_i(x_i - \overline{x})^2$$

The *standard deviation*, s, is defined as $\sqrt{s^2}$.

Define m_r as the rth moment about the mean $\overline{x}$:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} w_i(x_i - \overline{x})^r$$

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 pth *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 pth percentile is then

$$x_{[p]} = \begin{cases} \dfrac{x_{(i-1)} + x_{(i)}}{2} & \text{if } W_{(i-1)} = P \\ x_{(i)} & \text{otherwise} \end{cases}$$

References

Cox, N. J. 2010. Speaking Stata: The limits of sample skewness and kurtosis. *Stata Journal* 10: 482–495.

——. 2013. Speaking Stata: Trimming to taste. *Stata Journal* 13: 640–666.

David, H. A. 2001. First (?) occurrence of common terms in statistics and probability. In *Annotated Readings in the History of Statistics*, ed. H. A. David and A. W. F. Edwards, 209–246. New York: Springer.

Galton, F. 1875. Statistics by intercomparison, with remarks on the law of frequency of error. *Philosophical Magazine* 49: 33–46. https://doi.org/10.1080/14786447508641172.

Gelade, W., V. Verardi, and C. Vermandele. 2015. Time-efficient algorithms for robust estimators of location, scale, symmetry, and tail heaviness. *Stata Journal* 15: 77–94.

Gleason, J. R. 1997. sg67: Univariate summaries with boxplots. *Stata Technical Bulletin* 36: 23–25. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 179–183. College Station, TX: Stata Press.

——. 1999. sg67.1: Update to univar. *Stata Technical Bulletin* 51: 27–28. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 159–161. College Station, TX: Stata Press.

Hald, A. 1998. *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

Kirkwood, B. R., and J. A. C. Sterne. 2003. *Essential Medical Statistics*. 2nd ed. Malden, MA: Blackwell.

Lauritzen, S. L. 2002. *Thiele: Pioneer in Statistics*. Oxford: Oxford University Press.

Plackett, R. L. 1958. Studies in the history of probability and statistics: VII. The principle of the arithmetic mean. *Biometrika* 45: 130–135. https://doi.org/10.2307/2333051.

Pollock, P. H., III, and B. C. Edwards. 2019. *A Stata Companion to Political Analysis*. 4th ed. Thousand Oaks, CA: CQ Press.

Scott, L. J., and C. A. Rogers. 2015. Creating summary tables using the sumtable command. *Stata Journal* 15: 775–783.

Stigler, S. M. 1977. Fractional order statistics, with applications. *Journal of the American Statistical Association* 72: 544–550. https://doi.org/10.2307/2286215.

Stuart, A., and J. K. Ord. 1994. *Kendall's Advanced Theory of Statistics: Distribution Theory, Vol I*. 6th ed. London: Arnold.

Thiele, T. N. 1889. *Forelæsringer over Almindelig Iagttagelseslære: Sandsynlighedsregning og mindste Kvadraters Methode*. Kjøbenhavn: C.A. Reitzel. (English translation included in Lauritzen 2002).

Weisberg, H. F. 1992. *Central Tendency and Variability*. Newbury Park, CA: SAGE.

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** — Table of frequencies, summaries, and command results

[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