Title

stata.com

graph twoway histogram — Histogram plots

Description

`twoway histogram` draws histograms of *varname*. Also see [R] **histogram** for an easier-to-use alternative.

Quick start

Histogram of continuous variable `v1`

```
twoway histogram v1
```

Histogram of categorical variable `v2`

```
twoway histogram v2, discrete
```

As above, but place a gap between the bars by reducing bar width by 15%

```
twoway histogram v2, discrete gap(15)
```

As above, but with separate graph areas for each level of `catvar`

```
twoway histogram v2, discrete gap(15) by(catvar)
```

As above, and place graph areas in a single column

```
twoway histogram v2, discrete gap(15) by(catvar, cols(1))
```

Histogram of `v1` with bars scaled to reflect the number of observations in each bin

```
twoway histogram v1, frequency
```

As above, but with horizontal bars

```
twoway histogram v1, frequency horizontal
```

Histogram of `v1` with 10 bins

```
twoway histogram v1, bin(10)
```

Specify that the y axis should have markers and labels at 0, 25, 50, 75, and 100

```
twoway histogram v1, ylabel(0(25)100)
```

Menu

Graphics > Twoway graph (scatter, line, etc.)

Syntax

```
twoway histogram varname [if] [in] [weight]
    [ , [discrete_options | continuous_options] common_options ]
```

discrete_options	Description
`discrete`	specify that data are discrete
`width(#)`	width of bins in *varname* units
`start(#)`	theoretical minimum value

continuous_options	Description
`bin(#)`	# of bins
`width(#)`	width of bins in *varname* units
`start(#)`	lower limit of first bin

common_options	Description
`density`	draw as density; the default
`fraction`	draw as fractions
`frequency`	draw as frequencies
`percent`	draw as percents
`vertical`	vertical bars; the default
`horizontal`	horizontal bars
`gap(#)`	reduce width of bars, $0 \leq \# < 100$
barlook_options	change look of bars
axis_choice_options	associate plot with alternative axis
twoway_options	titles, legends, axes, added lines and text, by, regions, name, aspect ratio, etc.

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

Options for use in the discrete case

`discrete` specifies that *varname* is discrete and that each unique value of *varname* be given its own bin (bar of histogram).

`width(#)` is rarely specified in the discrete case; it specifies the width of the bins. The default is `width(`*d*`)`, where *d* is the observed minimum difference between the unique values of *varname*.

Specify `width()` if you are concerned that your data are sparse. For example, *varname* could in theory take on the values 1, 2, 3, ..., 9, but because of sparseness, perhaps only the values 2, 4, 7, and 8 are observed. Here the default width calculation would produce `width(2)`, and you would want to specify `width(1)`.

`start(#)` is also rarely specified in the discrete case; it specifies the theoretical minimum value of *varname*. The default is `start(`*m*`)`, where *m* is the observed minimum value.

As with `width()`, specify `start()` when you are concerned about sparseness. In the previous example, you would also want to specify `start(1)`. `start()` does nothing more than add white space to the left side of the graph.

`start()`, if specified, must be less than or equal to *m*, or an error will be issued.

Options for use in the continuous case

`bin(#)` and `width(#)` are alternatives that specify how the data are to be aggregated into bins. `bin()` specifies the number of bins (from which the width can be derived), and `width()` specifies the bin width (from which the number of bins can be derived).

If neither option is specified, the results are the same as if `bin(`*k*`)` were specified, where

$$k = \min\left(\sqrt{N}, 10 \times \frac{\ln(N)}{\ln(10)}\right)$$

and where *N* is the number of nonmissing observations of *varname*.

`start(#)` specifies the theoretical minimum of *varname*. The default is `start(`*m*`)`, where *m* is the observed minimum value of *varname*.

Specify `start()` when you are concerned about sparse data. For instance, you might know that *varname* can go down to 0, but you are concerned that 0 may not be observed.

`start()`, if specified, must be less than or equal to *m*, or an error will be issued.

Options for use in both cases

`density`, `fraction`, `frequency`, and `percent` are alternatives that specify whether you want the histogram scaled to density, fractional, or frequency units, or percentages. `density` is the default.

`density` scales the height of the bars so that the sum of their areas equals 1.

`fraction` scales the height of the bars so that the sum of their heights equals 1.

`frequency` scales the height of the bars so that each bar's height is equal to the number of observations in the category, and thus the sum of the heights is equal to the total number of nonmissing observations of *varname*.

`percent` scales the height of the bars so that the sum of their heights equals 100.

`vertical` and `horizontal` specify whether the bars are to be drawn vertically (the default) or horizontally.

`gap(#)` specifies that the bar width be reduced by # percent. `gap(0)` is the default; `histogram` sets the width so that adjacent bars just touch. If you wanted gaps between the bars, you would specify, for instance, `gap(5)`.

Also see [G-2] **graph twoway rbar** for other ways to set the display width of the bars. Histograms are actually drawn using `twoway rbar` with a restriction that 0 be included in the bars; `twoway histogram` will accept any options allowed by `twoway rbar`.

barlook_options set the look of the bars. The most important of these options is `color(`*colorstyle*`)`, which specifies the color of the bars; see [G-4] ***colorstyle*** for a list of color choices. See [G-3] ***barlook_options*** for information on the other *barlook_options*.

axis_choice_options associate the plot with a particular y or x axis on the graph; see [G-3] ***axis_choice_options***.

twoway_options are a set of common options supported by all `twoway` graphs. These options allow you to title graphs, name graphs, control axes and legends, add lines and text, set aspect ratios, create graphs over `by()` groups, and change some advanced settings. See [G-3] ***twoway_options***.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Relationship between graph twoway histogram and histogram

`graph twoway histogram`—documented here—and `histogram`—documented in [R] **histogram**—are almost the same command. `histogram` has the advantages that

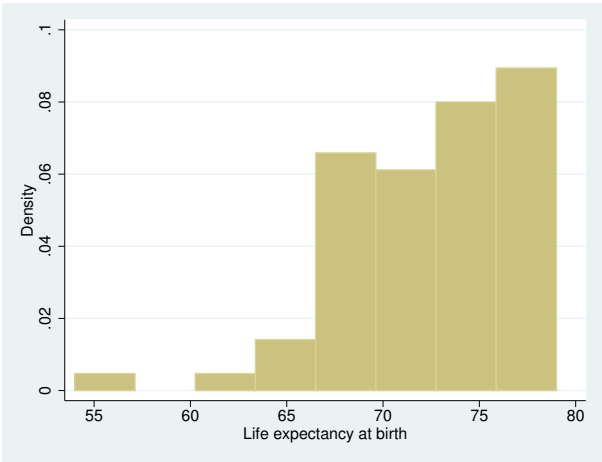
1. it allows overlaying of a normal density or a kernel estimate of the density;
2. if a density estimate is overlaid, it scales the density to reflect the scaling of the bars.

`histogram` is implemented in terms of `graph twoway histogram`.

Typical use

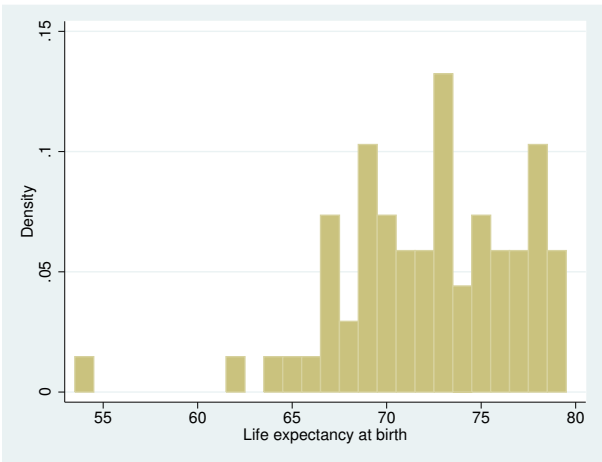
When you do not specify otherwise, `graph twoway histogram` assumes that the variable is continuous:

```
. use http://www.stata-press.com/data/r14/lifeexp
(Life expectancy, 1998)
. twoway histogram le
```



Even with a continuous variable, you may specify the `discrete` option to see the individual values:

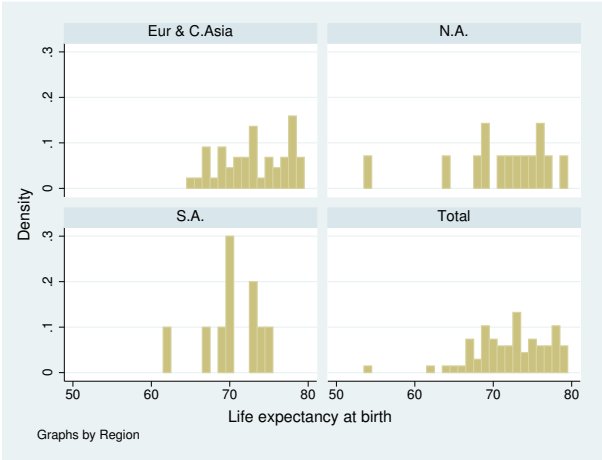
```
. twoway histogram le, discrete
```



Use with by()

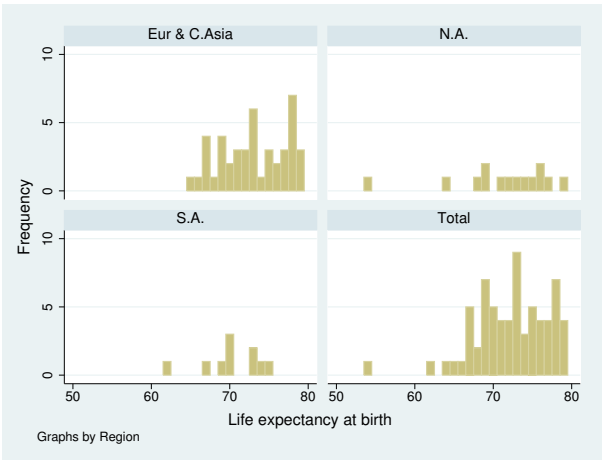
graph twoway histogram may be used with by():

```
. use http://www.stata-press.com/data/r14/lifeexp, clear
(Life expectancy, 1998)
. twoway histogram le, discrete by(region, total)
```



Here specifying frequency is a good way to show both the distribution and the overall contribution to the total:

```
. twoway histogram le, discrete freq by(region, total)
```



The height of the bars reflects the number of countries. Here—and in all the above examples—we would do better by obtaining population data on the countries and then typing

```
. twoway histogram le [fw=pop], discrete freq by(region, total)
```

so that bar height reflected total population.

History

According to Beniger and Robyn (1978, 4), although A. M. Guerry published a histogram in 1833, the word "histogram" was first used by Karl Pearson in 1895.

References

Beniger, J. R., and D. L. Robyn. 1978. Quantitative graphics in statistics: A brief history. *American Statistician* 32: 1–11.

Cox, N. J. 2005. Speaking Stata: Density probability plots. *Stata Journal* 5: 259–273.

——. 2007. Software Updates: Speaking Stata: Density probability plots. *Stata Journal* 7: 593.

Guerry, A.-M. 1833. *Essai sur la Statistique Morale de la France*. Paris: Crochard.

Harrison, D. A. 2005. Stata tip 20: Generating histogram bin variables. *Stata Journal* 5: 280–281.

Pearson, K. 1895. Contributions to the mathematical theory of evolution—II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London, Series A* 186: 343–414.

Also see

[R] **histogram** — Histograms for continuous and categorical variables

[G-2] **graph twoway kdensity** — Kernel density plots