lv — Letter-value displays

Description

`lv` shows a letter-value display (Tukey 1977, 44–49; Hoaglin 1983) for each variable in *varlist*. If no variables are specified, letter-value displays are shown for each numeric variable in the data.

Quick start

Letter-value display for all numeric variables in the dataset

```
lv
```

Letter-value display for `v1`

```
lv v1
```

Also generate new variables `_mid`, `_spread`, `_psigma`, and `_z2` containing midsummaries, spreads, pseudosigmas, and z^2 values

```
lv v1, generate
```

Letter-value displays for `v1` separately for each value of `v2`

```
bysort v2: lv v1
```

Menu

Statistics > Summaries, tables, and tests > Distributional plots and tests > Letter-value display

Syntax

`lv` [*varlist*] [*if*] [*in*] [`, generate tail(#)`]

`by` and `collect` are allowed; see **[U] 11.1.10 Prefix commands**.

Options

Main

`generate` adds four new variables to the data: `_mid`, containing the midsummaries; `_spread`, containing the spreads; `_psigma`, containing the pseudosigmas; and `_z2`, containing the squared values from a standard normal distribution corresponding to the particular letter value. If the variables `_mid`, `_spread`, `_psigma`, and `_z2` already exist, their contents are replaced. At most, only the first 11 observations of each variable are used; the remaining observations contain missing. If *varlist* specifies more than one variable, the newly created variables contain results for the last variable specified. The `generate` option may not be used with the `by` prefix.

`tail(#)` indicates the inverse of the tail density through which letter values are to be displayed: 2 corresponds to the median (meaning half in each tail), 4 to the fourths (roughly the 25th and 75th percentiles), 8 to the eighths, and so on. # may be specified as 4, 8, 16, 32, 64, 128, 256, 512, or 1,024 and defaults to a value of # that has corresponding depth just greater than 1. The default is taken as 1,024 if the calculation results in a number larger than 1,024. Given the intelligent default, this option is rarely specified.

Remarks and examples

Letter-value displays are a collection of observations drawn systematically from the data, focusing especially on the tails rather than the middle of the distribution. The displays are called letter-value displays because letters have been (almost arbitrarily) assigned to tail densities:

Letter	Tail area	Letter	Tail area
M	1/2	B	1/64
F	1/4	A	1/128
E	1/8	Z	1/256
D	1/16	Y	1/512
C	1/32	X	1/1024

▷ Example 1

We have data on the mileage ratings of 74 automobiles. To obtain a letter-value display, we type

```
. use https://www.stata-press.com/data/r19/auto
(1978 automobile data)

. lv mpg

 #     74                Mileage (mpg)
                 _________________________________

 M     37.5   |                20                 |       spread  pseudosigma
 F     19     |       18       21.5          25   |            7     5.216359
 E     10     |       15       21.5          28   |           13     5.771728
 D      5.5   |       14       22.25         30.5 |         16.5     5.576303
 C      3     |       14       24.5          35   |           21     5.831039
 B      2     |       12       23.5          35   |           23     5.732448
 A      1.5   |       12       25            38   |           26     6.040635
        1     |       12       26.5          41   |           29      6.16562
              |                                   |
              |                                   |      # below      # above
inner fence   |      7.5                     35.5 |            0            1
outer fence   |       -3                     46   |            0            0
```

The decimal points can be made to line up and thus the output made more readable by specifying a display format for the variable; see [U] 12.5 Formats: Controlling how data are displayed.

```
. format mpg %9.2f
. lv mpg
  #      74              Mileage (mpg)
                 _________________________________
  M      37.5  |               20.00               |      spread  pseudosigma
  F      19    |     18.00     21.50     25.00     |        7.00         5.22
  E      10    |     15.00     21.50     28.00     |       13.00         5.77
  D       5.5  |     14.00     22.25     30.50     |       16.50         5.58
  C       3    |     14.00     24.50     35.00     |       21.00         5.83
  B       2    |     12.00     23.50     35.00     |       23.00         5.73
  A       1.5  |     12.00     25.00     38.00     |       26.00         6.04
          1    |     12.00     26.50     41.00     |       29.00         6.17
               |                                   |
               |                                   |     # below      # above
inner fence    |      7.50               35.50     |           0            1
outer fence    |     -3.00               46.00     |           0            0
```

At the top, the number of observations is indicated as 74. The first line shows the statistics associated with M, the letter value that puts half the density in each tail, or the median. The median has *depth* 37.5 (that is, in the ordered data, M is 37.5 observations in from the extremes) and has value 20. The next line shows the statistics associated with F or the fourths. The fourths have depth 19 (that is, in the ordered data, the lower fourth is observation 19, and the upper fourth is observation $74 - 19 + 1$), and the values of the lower and upper fourths are 18 and 25. The number in the middle is the point halfway between the fourths—called a midsummary. If the distribution were perfectly symmetric, the midsummary would equal the median. The spread is the difference between the lower and upper summaries ($25 - 18 = 7$). For fourths, half the data lie within a 7-mpg band. The pseudosigma is a calculation of the standard deviation using only the lower and upper summaries and assuming that the variable is normally distributed. If the data really were normally distributed, all the pseudosigmas would be roughly equal.

After the letter values, the line labeled with depth 1 reports the minimum and maximum values. Here the halfway point between the extremes is 26.5, which is greater than the median, indicating that 41 is more extreme than 12, at least relative to the median. And with each letter value, the midsummaries are increasing—our data are skewed. The pseudosigmas are also increasing, indicating that the data are spreading out relative to a normal distribution, although, given the evident skewness, this elongation may be an artifact of the skewness.

At the end is an attempt to identify outliers, although the points so identified are merely outside some predetermined cutoff. Points outside the inner fence are called *outside values* or *mild outliers*. Points outside the outer fence are called *severe outliers*. The inner fence is defined as $(3/2)$IQR and the outer fence as 3IQR above and below the F summaries, where the interquartile range (IQR) is the spread of the fourths.

◁

❑ Technical note

The form of the letter-value display has varied slightly with different authors. `lv` displays appear as described by Hoaglin (1983) but as modified by Emerson and Stoto (1983), where they included the midpoint of each of the spreads. This format was later adopted by Hoaglin (1985). If the distribution is symmetric, the midpoints will all be roughly equal. On the other hand, if the midpoints vary systematically, the distribution is skewed.

The pseudosigmas are obtained from the lower and upper summaries for each letter value. For each letter value, they are the standard deviation a normal distribution would have if its spread for the given letter value were to equal the observed spread. If the pseudosigmas are all roughly equal, the data are said to have *neutral elongation*. If the pseudosigmas increase systematically, the data are said to be more elongated than a normal, that is, have thicker tails. If the pseudosigmas decrease systematically, the data are said to be less elongated than a normal, that is, have thinner tails.

Interpretation of the number of mild and severe outliers is more problematic. The following discussion is drawn from Hamilton (1991):

Obviously, the presence of any such outliers does not rule out that the data have been drawn from a normal distribution; in large datasets, there will most certainly be observations outside $(3/2)$IQR and 3IQR. Severe outliers, however, make up about two per million (0.0002%) of a normal population. In samples, they lie far enough out to have substantial effects on means, standard deviations, and other classical statistics. The 0.0002%, however, should be interpreted carefully; outliers appear more often in small samples than one might expect from population proportions because of sampling variation in estimated quartiles. Monte Carlo simulation by Hoaglin, Iglewicz, and Tukey (1986) obtained these results on the percentages and numbers of outliers in random samples from a normal population:

	Percentage		Number	
n	any outliers	severe	any outliers	severe
10	2.83	0.362	0.283	0.0362
20	1.66	0.074	0.332	0.0148
50	1.15	0.011	0.575	0.0055
100	0.95	0.002	0.95	0.002
200	0.79	0.001	1.58	0.002
300	0.75	0.001	2.25	0.003
∞	0.70	0.0002	∞	∞

Thus, the presence of any severe outliers in samples of less than 300 is sufficient to reject normality. Hoaglin, Iglewicz, and Tukey (1981) suggested the approximation $0.00698 + 0.4/n$ for the fraction of mild outliers in a sample of size n or, equivalently, $0.00698n + 0.4$ for the number of outliers.

❑

Example 2

The generate option adds the _mid, _spread, _psigma, and _z2 variables to our data, making possible many of the diagnostic graphs suggested by Hoaglin (1985).

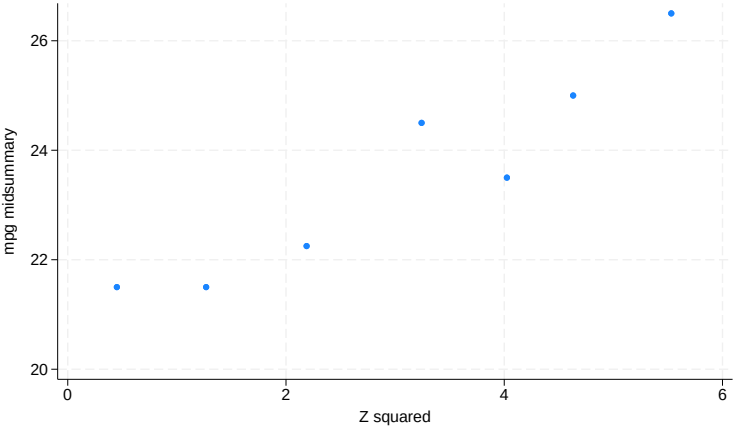
```
. lv mpg, generate
  (output omitted)
. list _mid _spread _psigma _z2 in 1/12
```

	_mid	_spread	_psigma	_z2
1.	20	.	.	.
2.	21.5	7	5.216359	.4501955
3.	21.5	13	5.771728	1.26828
4.	22.25	16.5	5.576303	2.188846
5.	24.5	21	5.831039	3.24255
6.	23.5	23	5.732448	4.024532
7.	25	26	6.040635	4.631499
8.	.	.	.	.
9.	.	.	.	.
10.	.	.	.	.
11.	26.5	29	6.16562	5.53073
12.	.	.	.	.

Observations 12 through the end are missing for these new variables. The definition of the observations is always the same. The first observation contains the M summary; the second, the F; the third, the E; and so on. Observation 11 always contains the summary for depth 1. Observations 8–10—corresponding to letter values Z, Y, and X—contain missing because these statistics were not calculated. We have only 74 observations, and their depth would be 1.

Hoaglin (1985) suggests graphing the midsummary against z^2. If the distribution is not skewed, the points in the resulting graph will be along a horizontal line:

```
. scatter _mid _z2
```



The graph clearly indicates the skewness of the distribution. We might also graph _psigma against _z2 to examine elongation.

◁

Stored results

lv stores the following in r():

Scalars

r(N)	number of observations	r(u_C)	upper 32nd
r(min)	minimum	r(l_B)	lower 64th
r(max)	maximum	r(u_B)	upper 64th
r(median)	median	r(l_A)	lower 128th
r(l_F)	lower 4th	r(u_A)	upper 128th
r(u_F)	upper 4th	r(l_Z)	lower 256th
r(l_E)	lower 8th	r(u_Z)	upper 256th
r(u_E)	upper 8th	r(l_Y)	lower 512th
r(l_D)	lower 16th	r(u_Y)	upper 512th
r(u_D)	upper 16th	r(l_X)	lower 1024th
r(l_C)	lower 32nd	r(u_X)	upper 1024th

The lower/upper 8ths, 16ths, …, 1024ths will be defined only if there are sufficient data.

Methods and formulas

Let N be the number of (nonmissing) observations on x, and let $x_{(i)}$ refer to the ordered data when i is an integer. Define $x_{(i+0.5)} = (x_{(i)} + x_{(i+1)})/2$; the median is defined as $x_{\{(N+1)/2\}}$.

Define $x_{[d]}$ as the pair of numbers $x_{(d)}$ and $x_{(N+1-d)}$, where d is called the *depth*. Thus, $x_{[1]}$ refers to the minimum and maximum of the data. Define $m = (N+1)/2$ as the depth of the median, $f = (\lfloor m \rfloor + 1)/2$ as the depth of the fourths, $e = (\lfloor f \rfloor + 1)/2$ as the depth of the eighths, and so on. Depths are reported on the far left of the letter-value display. The corresponding fourths of the data are $x_{[f]}$, the eighths are $x_{[e]}$, and so on. These values are reported inside the display. The middle value is defined as the corresponding midpoint of $x_{[\cdot]}$. The spreads are defined as the difference in $x_{[\cdot]}$.

The corresponding point z_i on a standard normal distribution is obtained as (Hoaglin 1985, 456–457)

$$z_i = \begin{cases} F^{-1}\{(d_i - 1/3)/(N + 1/3)\} & \text{if} d_i > 1 \\ F^{-1}\{0.695/(N + 0.390)\} & \text{otherwise} \end{cases}$$

where d_i is the depth of the letter value. The corresponding pseudosigma is obtained as the ratio of the spread to $-2z_i$ (Hoaglin 1985, 431).

Define $(F_l, F_u) = x_{[f]}$. The inner fence has cutoffs $F_l - \frac{3}{2}(F_u - F_l)$ and $F_u + \frac{3}{2}(F_u - F_l)$. The outer fence has cutoffs $F_l - 3(F_u - F_l)$ and $F_u + 3(F_u - F_l)$.

The inner-fence values reported by lv are almost equal to those used by graph, box to identify outside points. The only difference is that graph uses a slightly different definition of fourths, namely, the 25th and 75th percentiles as defined by summarize; see [R] **summarize**.

References

Cox, N. J. 2016. Speaking Stata: Letter values as selected quantiles. *Stata Journal* 16: 1058–1071.

Emerson, J. D., and M. A. Stoto. 1983. "Transforming data". In *Understanding Robust and Exploratory Data Analysis*, edited by D. C. Hoaglin, C. F. Mosteller, and J. W. Tukey, 97–128. New York: Wiley.

Fox, J. 1990. "Describing univariate distributions". In *Modern Methods of Data Analysis*, edited by J. Fox and J. S. Long, 58–125. Newbury Park, CA: Sage.

Hamilton, L. C. 1991. sed4: Resistant normality check and outlier identification. *Stata Technical Bulletin* 3: 15–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 86–90. College Station, TX: Stata Press.

Hoaglin, D. C. 1983. "Letter values: A set of selected order statistics". In *Understanding Robust and Exploratory Data Analysis*, edited by D. C. Hoaglin, C. F. Mosteller, and J. W. Tukey, 33–57. New York: Wiley.

———. 1985. "Using quantiles to study shape". In *Exploring Data Tables, Trends, and Shapes*, edited by D. C. Hoaglin, C. F. Mosteller, and J. W. Tukey, 417–460. New York: Wiley.

Hoaglin, D. C., B. Iglewicz, and J. W. Tukey. 1981. "Small-sample performance of a resistant rule for outlier detection". In *1980 Proceedings of the Statistical Computing Section*. Washington, DC: American Statistical Association.

———. 1986. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association* 81: 991–999. https://doi.org/10.2307/2289073.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

Also see

[R] **Diagnostic plots** — Distributional diagnostic plots

[R] **stem** — Stem-and-leaf displays

[R] **summarize** — Summary statistics