NAG Toolbox for MATLAB

g01ar

1 Purpose

g01ar produces a stem and leaf display for a single sample of observations.

2 Syntax

```
[unit, plot, lines, sorty, ifail] = g01ar(range, prt, y, nstepx, nstepy,
unit, 'n', n)
```

3 Description

g01ar produces a stem and leaf display for a single sample of n observations. The stem and leaf display shows data values separated into the form of a 'stem' and a 'leaf'. For example, a value of 473 could be represented as 47 3 where the stem is 47 and the leaf is 3. The data is scaled using a value known as the 'leaf digit unit'. In the above example the leaf digit unit would be 1.0.

The following example illustrates a stem and leaf display.

For the 10 observations:

1.8 2.3 2.1 1.9 2.1 2.4 2.0 2.0 1.9 2.1

the stem and leaf display is:

```
1  1  8
3  1  99
5  2  00
5  2  111
2  2
2  2  3
1  2  4
```

where the leaf digit unit is 0.1 so that 1 8 represents 1.8 (i.e., 18×0.1). The leaf digit unit distinguishes between the numbers 18.0, 1.8, 0.18, etc. which may otherwise all be represented by 1 8.

Included in the above display is an initial column specifying the cumulative count of values, up to and including that particular line, from either the top or bottom of the display, whichever is smaller. An exception to this is when the line on which the median lies is reached, in which case the actual count of values on that line is displayed, rather than a cumulative count, and this is highlighted by enclosing the count in parentheses. In this case the median is 2.05 and thus falls between the two lines at which the cumulative count has reached $n/2$ where n is the number of observations.

Some of the other features of the stem and leaf display are illustrated by the following two examples.

For the 30 observations:

−19.0	−3.0	−1.0	0.0	1.0	2.0	2.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0	4.0	5.0	5.0	5.0	5.0	6.0
6.0	6.0	7.0	7.0	8.0	10.0	11.0	11.0	13.0	31.0

the stem and leaf display may be:

```
 1   1.  9
 1   1*
 1  -0.
 3  -0*  13
15  +0*  012233344444
15  +0.  55556667788
 5   1*  011
 2   1.  3
 1   2
 1   2.
```

```
 1   3   1
```

In the above display all the data are plotted and the leaf digit unit is 1.0. Also in this display different leaves, that is different digits, may be plotted on a particular line. In this case we have 5 possible digits per line, that is 2 lines per stem, and these are represented as follows:

`*` indicates that the line may contain the digits 0 to 4;

`.` indicates that the line may contain the digits 5 to 9.

Alternatively the stem and leaf display may look like:

```
      LO   -19

  2   -0*  3
  3   +0T  1
  5   +0*  01
 10   +0T  22333
( 9)  +0F  444445555
 11   +0*  66677
  6   +0T  8
  5    1*  011
  2    1T  3

      HI   31
```

Again the leaf digit unit is 1.0 but in this display just the data between the fences, which are the hinges $\pm 1\frac{1}{2}\times$ the inter-hinge range, are plotted. Any data points that fall outside the fences are presented separately in the display under the headings `LO` for those points below the lower fence and `HI` for those points above the upper fence.

Again in this display different leaves, that is different digits, may be plotted on a particular line. However in this case we have 2 possible digits per line, that is 5 lines per stem, and these are represented as follows

`*` indicates that the line may contain the digits 0 or 1;

`T` indicates that the line may contain the digits 2 or 3;

`F` indicates that the line may contain the digits 4 or 5;

`S` indicates that the line may contain the digits 6 or 7;

`.` indicates that the line may contain the digits 8 or 9.

A display may also allow 10 different digits (0 to 9) per line, that is 1 line per stem, or just 1 digit per line, that is 10 lines per stem, as in the first of the three examples above.

Note that the median here is 4.5. This falls between two lines in the first display but is highlighted on the second display since it lies on a particular line.

Finally if there are positive and negative numbers on the display these are highlighted by a $+$ or $-$ sign where the distinction is required, that is near the zero-point.

If there are too many leaves to fit in the plot width allowed, g01ar plots as many leaves as possible and places an asterisk to the right to indicate that some leaves are not displayed. If this occurs and you wish to be able to plot all the leaves then the width of the plot may be adjusted.

Options also allow the leaf unit and the height of the display to be specified by you or calculated by g01ar. These parameters may be used to control the type of the display you wish to obtain. Fixing the unit and changing the height of the display may alter the number of lines used per stem, that is the number of different digits per line. g01ar will choose a display for the fixed unit that attempts to make as much use of the available height as possible, thus increasing the height may allow for more lines per stem whereas decreasing the height may force the display to use fewer lines per stem. Similarly you may wish to fix the height and vary the leaf digit unit used on the display. See Section 8 for further details.

The display is returned in a character array with the option of printing the display.

4 References

Erickson B H and Nosanchuk T A 1985 *Understanding Data* Open University Press, Milton Keynes

Tukey J W 1977 *Exploratory Data Analysis* Addison–Wesley

Velleman P F and Hoaglin D C 1981 *Applications, Basics, and Computing of Exploratory Data Analysis* Duxbury Press, Boston, MA

5 Parameters

5.1 Compulsory Input Parameters

1: **range – string**

Indicates whether you wish to scale the plot to the extremes of the data or to the fences.

range = 'E'

The display is a plot to the extremes, that is a plot of all the data.

range = 'F'

The display is a plot of the data between the fences.

Constraint: **range** must be one of 'E' or 'F'.

2: **prt – string**

Indicates whether the stem and leaf display is to be output to an external file.

prt = 'N'

The display is not output to an external file.

prt = 'P'

The display is output to the current advisory message unit as defined by x04ab. Only the first 132 characters of each line are actually printed.

Constraint: **prt** must be one of 'P' or 'N'.

3: **y(n) – double array**

The n observations.

4: **nstepx – int32 scalar**

the number of character positions to be plotted horizontally.

Constraint: **nstepx** ≥ 35.

5: **nstepy – int32 scalar**

The maximum number of character positions to be plotted vertically.

If **nstepy** ≤ 0 a suitable value will be used by g01ar for the number of character positions to be plotted vertically. This will clearly be less than or equal to the value of **ldplot**.

Constraint: **nstepy** ≤ 0 or **nstepy** ≥ 5.

6: **unit – double scalar**

Indicates the leaf digit unit to be used.

If **unit** > 0.0 and is not a power of ten, it will be converted to the nearest power of ten below the input value for unit.

If **unit** ≤ 0.0, the optimum unit will be used. This is based on the range of the data to be plotted and the number of lines available for the display.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **y**, **sorty**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: $\mathbf{n} \geq 2$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldplot, iwork

5.4 Output Parameters

1: **unit – double scalar**

Contains the actual unit used in the stem and leaf display.

2: **plot(ldplot,nstepx) – string array**

The stem and leaf display.

3: **lines – int32 scalar**

The actual number of lines needed for the display.

4: **sorty(n) – double array**

The observations sorted into ascending order.

5: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, $\mathbf{n} < 2$,
or $\mathbf{nstepx} < 35$,
or $0 < \mathbf{nstepy} < 5$,
or $\mathbf{ldplot} < 5$,
or $\mathbf{ldplot} < \mathbf{nstepy}$.

ifail $= 2$

On entry, **prt** $\neq$ 'P' or 'N',
or **range** $\neq$ 'E' or 'F'.

ifail $= 3$

The number of lines needed to produce the display exceeds the maximum number of lines allowed. You may wish to increase **nstepy**.

ifail $= 4$

One of the observations is too large and causes a value to exceed the maximum integer allowed.

7 Accuracy

Accuracy is limited by the number of significant figures that may be represented on the display which will depend on the data, the number of lines available and the unit used.

8 Further Comments

g01ar uses integer representations of the data. If very large data values are being used they should be scaled before using this function. The largest integer can be found by calling x02bb.

If an asterisk is plotted at the end of a line to indicate that some leaves are not displayed you should increase **nstepx** if they wish to be able to print the rest of the leaves on that line.

Note that if you request g01ar to print the plot only the first 132 characters of each line are printed. The full plot is stored in the array **plot** so you do have the option of printing a plot which has more than 132 characters on a line.

When the leaf digit unit is set, the number of lines per stem is decided as follows:

Let r be the range of the data to be plotted:

r = largest observation – smallest observation: if all the data to both extremes are to be plotted (that is if **range** = 'E'),

r = upper fence – lower fence: if only the data between the fences are to be plotted (that is if **range** = 'F').

Let l be the number of lines available for the plot:

$l =$ **nstepy** $- 4$ if **nstepy** > 0,

$l =$ **ldplot** $- 4$ if **nstepy** ≤ 0.

The 4 lines are subtracted to allow space for the display headings. If only the data between the fences are to be plotted then l must be further reduced to allow space to present those values outside the fences. This will involve a minimum of another 4 lines.

Let $e = \dfrac{(r/\textbf{unit}) + 1}{l}$,

then the number of lines per stem is:

$$\begin{aligned} 1 &\text{ if } 5 < e \leq 10, \text{ that is digits per line is 10,} \\ 2 &\text{ if } 2 < e \leq 5, \text{ that is digits per line is 5,} \\ 5 &\text{ if } 1 < e \leq 2, \text{ that is digits per line is 2,} \\ 10 &\text{ if } 0 < e \leq 1, \text{ that is digits per line is 1.} \end{aligned}$$

The time taken by the function increases with n.

9 Example

```
range = 'Fences';
prt = 'Print';
y = [31;
     1;
     2;
     3;
     4;
     5;
     6;
     7;
     8;
     -9;
     1;
     2;
     3;
```

```
     4;
     5;
     6;
     7;
     8;
     2;
     3;
     4;
     5;
     6;
     7;
     3;
     4;
     5;
     6;
     4;
     5];
nstepx = int32(72);
nstepy = int32(20);
unit = 0;
[unitOut, plot, lines, sorty, ifail] = g01ar(range, prt, y, nstepx,
nstepy, unit)
```

```
Stem-and-leaf display
 Leaf digit unit = 1.0
 1  2  represents  12.

       LO  -9

   3    0  11
   6    0  222
  10    0  3333
  15    0  44444
  15    0  55555
  10    0  6666
   6    0  777
   3    0  88

       HI  31
unitOut =
     1
plot =
     array elided
lines =
          16
sorty =
    -9
     1
     1
     2
     2
     2
     3
     3
     3
     3
     4
     4
     4
     4
     4
     5
     5
     5
     5
     5
     6
     6
     6
     6
```

```
      7
      7
      7
      8
      8
     31
ifail =
           0
```