NAG Toolbox for MATLAB

g02bh

1 Purpose

g02bh computes means and standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for selected variables omitting completely any cases with a missing observation for any variable (either over all variables in the data set or over only those variables in the selected subset).

2 Syntax

```
[miss, xmiss, xbar, std, ssp, r, ncases, ifail] = g02bh(n, x, miss,
xmiss, mistyp, kvar, 'm', m, 'nvars', nvars)
```

3 Description

The input data consists of n observations for each of m variables, given as an array

$$\left[x_{ij}\right], \qquad i = 1,2,\ldots,n(n \geq 2), j = 1,2,\ldots,m(m \geq 2),$$

where x_{ij} is the ith observation on the jth variable, together with the subset of these variables, $v_1, v_2, \ldots, v_p$, for which information is required.

In addition, each of the m variables may optionally have associated with it a value which is to be considered as representing a missing observation for that variable; the missing value for the jth variable is denoted by xm_j. Missing values need not be specified for all variables. The missing values can be utilized in two slightly different ways; you can indicate which scheme is required.

Firstly, let $w_i = 0$ if observation i contains a missing value for any of those variables in the set $1,2,\ldots,m$ for which missing values have been declared, i.e., if $x_{ij} = xm_j$ for any j ($j = 1,2,\ldots,m$) for which an xm_j has been assigned (see also Section 7); and $w_i = 1$ otherwise, for $i = 1,2,\ldots,n$.

Secondly, let $w_i = 0$ if observation i contains a missing value for any of those variables in the selected subset $v_1, v_2, \ldots, v_p$ for which missing values have been declared, i.e., if $x_{ij} = xm_j$ for any j ($j = v_1, v_2, \ldots, v_p$) for which an xm_j has been assigned (see also Section 7); and $w_i = 1$ otherwise, for $i = 1,2,\ldots,n$.

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i}, \qquad j = v_1, v_2, \ldots, v_p.$$

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(x_{ij} - \bar{x}_j\right)^2}{\sum_{i=1}^{n} w_i - 1}}, \qquad j = v_1, v_2, \ldots, v_p.$$

(c) Sums of squares and cross-products of deviations from means:

$$S_{jk} = \sum_{i=1}^{n} w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad j,k = v_1, v_2, \ldots, v_p.$$

(d) Pearson product-moment correlation coefficients:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}, \qquad j,k = v_1, v_2, \ldots, v_p.$$

If S_{jj} or S_{kk} is zero, R_{jk} is set to zero.

4 References

None.

5 Parameters

5.1 Compulsory Input Parameters

1: **n – int32 scalar**

n, the number of observations or cases.

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

2: **x(ldx,m) – double array**

ldx, the first dimension of the array, must be at least **n**.

$\mathbf{x}(i,j)$ must be set to x_{ij}, the value of the ith observation on the jth variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

3: **miss(m) – int32 array**

$\mathbf{miss}(j)$ must be set equal to 1 if a missing value, xm_j, is to be specified for the jth variable in the array **x**, or set equal to 0 otherwise. Values of **miss** must be given for all m variables in the array **x**.

4: **xmiss(m) – double array**

$\mathbf{xmiss}(j)$ must be set to the missing value, xm_j, to be associated with the jth variable in the array **x**, for those variables for which missing values are specified by means of the array **miss** (see Section 7).

5: **mistyp – int32 scalar**

Indicates the manner in which missing observations are to be treated.

mistyp $= 1$

A case is excluded if it contains a missing value for any of the variables $1, 2, \ldots, m$.

mistyp $= 0$

A case is excluded if it contains a missing value for any of the $p (\leq m)$ variables specified in the array **kvar**.

6: **kvar(nvars) – int32 array**

$\mathbf{kvar}(j)$ must be set to the column number in **x** of the jth variable for which information is required, for $j = 1, 2, \ldots, p$.

Constraint: $1 \leq \mathbf{kvar}(j) \leq \mathbf{m}$, for $j = 1, 2, \ldots, p$.

5.2 Optional Input Parameters

1: **m – int32 scalar**

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

m, the number of variables.

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

2: **nvars – int32 scalar**

Default: The dimension of the arrays **xbar**, **std**, **ssp**, **r**. (An error is raised if these dimensions are not equal.)

p, the number of variables for which information is required.

Constraint: $2 \leq \mathbf{nvars} \leq \mathbf{m}$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldssp, ldr

5.4 Output Parameters

1: **miss(m) – int32 array**

The array **miss** contains the function, and the information it contained on entry is lost.

2: **xmiss(m) – double array**

The array **xmiss** contains the function, and the information it contained on entry is lost.

3: **xbar(nvars) – double array**

The mean value, of $\bar{x}_j$, of the variable specified in $\mathbf{kvar}(j)$, for $j = 1, 2, \ldots, p$.

4: **std(nvars) – double array**

The standard deviation, s_j, of the variable specified in $\mathbf{kvar}(j)$, for $j = 1, 2, \ldots, p$.

5: **ssp(ldssp,nvars) – double array**

$\mathbf{ssp}(j,k)$ is the cross-product of deviations, S_{jk}, for the variables specified in $\mathbf{kvar}(j)$ and $\mathbf{kvar}(k)$, for $j, k = 1, 2, \ldots, p$.

6: **r(ldr,nvars) – double array**

$\mathbf{r}(j,k)$ is the product-moment correlation coefficient, R_{jk}, between the variables specified in $\mathbf{kvar}(j)$ and $\mathbf{kvar}(k)$, for $j, k = 1, 2, \ldots, p$.

7: **ncases – int32 scalar**

The number of cases actually used in the calculations (when cases involving missing values have been eliminated).

8: **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, **n** < 2.

ifail $= 2$

On entry, **nvars** < 2,
or **nvars** $>$ **m**.

ifail $= 3$

On entry, **ldx** $<$ **n**,
or **ldssp** $<$ **nvars**,
or **ldr** $<$ **nvars**.

ifail $= 4$

On entry, **kvar**$(j) < 1$,
or **kvar**$(j) >$ **m** for some $j = 1, 2, \ldots,$ **nvars**.

ifail $= 5$

On entry, **mistyp** $\neq 1$ or 0

ifail $= 6$

After observations with missing values were omitted, no cases remained.

ifail $= 7$

After observations with missing values were omitted, only one case remained.

7 Accuracy

g02bh does not use ***additional precision*** arithmetic for the accumulation of scalar products, so there may be a loss of significant figures for large n.

You are warned of the need to exercise extreme care in your selection of missing values. g02bh treats all values in the inclusive range $(1 \pm \text{ACC}) \times xm_j$, where xm_j is the missing value for variable j specified by you, and ACC is a machine-dependent constant as missing values for variable j.

You must therefore ensure that the missing value chosen for each variable is sufficiently different from all valid values for that variable so that none of the valid values fall within the range indicated above.

8 Further Comments

The time taken by g02bh depends on n and p, and the occurrence of missing values.

The function uses a two-pass algorithm.

9 Example

```
n = int32(5);
x = [3, 3, 1, 2;
     6, 4, -1, 4;
     9, 0, 5, 9;
     12, 2, 0, 0;
     -1, 5, 4, 12];
```

```
miss = [int32(0);
     int32(1);
     int32(0);
     int32(1)];
xmiss = [0; 0; 0; 0];
mistyp = int32(0);
kvar = [int32(4);
     int32(1);
     int32(2)];
[missOut, xmissOut, xbar, std, ssp, r, ncases, ifail] = g02bh(n, x, miss,
xmiss, mistyp, kvar)
```

```
missOut =
           4
           1
           2
           1
xmissOut =
     0
     0
     0
     0
xbar =
    6.0000
    2.6667
    4.0000
std =
    5.2915
    3.5119
    1.0000
ssp =
   56.0000  -30.0000   10.0000
  -30.0000   24.6667   -4.0000
   10.0000   -4.0000    2.0000
r =
    1.0000   -0.8072    0.9449
   -0.8072    1.0000   -0.5695
    0.9449   -0.5695    1.0000
ncases =
           3
ifail =
           0
```