

NAG Toolbox for MATLAB

g02be

1 Purpose

g02be computes means and standard deviations of variables, sums of squares and cross-products about zero, and correlation-like coefficients for a set of data omitting completely any cases with a missing observation for any variable.

2 Syntax

```
[miss, xmiss, xbar, std, sspz, rz, ncases, ifail] = g02be(n, x, miss,
xmiss, 'm', m)
```

3 Description

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

$$[x_{ij}], \quad i = 1, 2, \dots, n(n \geq 2), j = 1, 2, \dots, m(m \geq 2),$$

where x_{ij} is the i th observation on the j th variable. 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 j th variable is denoted by xm_j . Missing values need not be specified for all variables.

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

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^n w_i x_{ij}}{\sum_{i=1}^n w_i}, \quad j = 1, 2, \dots, m.$$

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^n w_i (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^n w_i - 1}}, \quad j = 1, 2, \dots, m.$$

(c) Sums of squares and cross-products about zero:

$$\tilde{S}_{jk} = \sum_{i=1}^n w_i x_{ij} x_{ik}, \quad j, k = 1, 2, \dots, m.$$

(d) Correlation-like coefficients:

$$\tilde{R}_{jk} = \frac{\tilde{S}_{jk}}{\sqrt{\tilde{S}_{jj} \tilde{S}_{kk}}}, \quad j, k = 1, 2, \dots, m.$$

If \tilde{S}_{jj} or \tilde{S}_{kk} is zero, \tilde{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: $n \geq 2$.
- 2: **x(ldx,m)** – **double** array
ldx, the first dimension of the array, must be at least **n**.
x(i,j) must be set to x_{ij} , the value of the i th observation on the j th variable, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$.
- 3: **miss(m)** – **int32** array
mmiss(j) must be set equal to 1 if a missing value, x_{mj} , is to be specified for the j th 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
xmiss(j) must be set to the missing value, x_{mj} , to be associated with the j th variable in the array **x**, for those variables for which missing values are specified by means of the array **miss** (see Section 7).

5.2 Optional Input Parameters

- 1: **m** – **int32** scalar
Default: The dimension of the arrays **x**, **xbar**, **std**, **ssp**, **r**. (An error is raised if these dimensions are not equal.)
 m , the number of variables.
Constraint: $m \geq 2$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldsspz, ldrz

5.4 Output Parameters

- 1: **miss(m)** – **int32** array
The array **miss** contains the function, and 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(m)** – **double** array
The mean value, \bar{x}_j , of the j th variable, for $j = 1, 2, \dots, m$.
- 4: **std(m)** – **double** array
The standard deviation, s_j , of the j th variable, for $j = 1, 2, \dots, m$.

- 5: **sspz(ldsspz,m) – double array**
sspz(j,k) is the cross-product about zero, \tilde{S}_{jk} , for $j, k = 1, 2, \dots, m$.
- 6: **rz(ldrz,m) – double array**
rz(j,k) is the correlation-like coefficient, \tilde{R}_{jk} , between the j th and k th variables, for $j, k = 1, 2, \dots, m$.
- 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, **m** < 2.

ifail = 3

On entry, **ldx** < **n**,
or **ldsspz** < **m**,
or **ldrz** < **m**.

ifail = 4

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

ifail = 5

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

7 Accuracy

g02be 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. g02be 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 g02be depends on n and m , and the occurrence of missing values.

The function uses a two-pass algorithm.

9 Example

```

n = int32(5);
x = [2, 3, 3;
     4, 6, 4;
     9, 9, 0;
     0, 12, 2;
     12, -1, 5];
miss = [int32(1);
        int32(0);
        int32(1)];
xmiss = [0;
         0;
         0];
[missOut, xmissOut, xbar, std, sspz, rz, ncases, ifail] = g02be(n, x,
miss, xmiss)

missOut =
         1
         3
         1
xmissOut =
         0
         0
         0
xbar =
    6.0000
    2.6667
    4.0000
std =
    5.2915
    3.5119
    1.0000
sspz =
    164     18     82
     18     46     28
     82     28     50
rz =
    1.0000    0.2072    0.9055
    0.2072    1.0000    0.5838
    0.9055    0.5838    1.0000
ncases =
         3
ifail =
         0

```