NAG Toolbox for MATLAB

g02bf

1 Purpose

g02bf 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 cases with missing values from only those calculations involving the variables for which the values are missing.

2 Syntax

```
[xbar, std, sspz, rz, ncases, cnt, ifail] = g02bf(n, x, miss, xmiss,
'm', m)
```

3 Description

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

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

where x_{ij} is the ith observation on the jth variable. 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.

Let $w_{ij} = 0$ if the ith observation for the jth variable is a missing value, i.e., if a missing value, xm_j, has been declared for the jth variable, and $x_{ij} = xm_j$ (see also Section 7); and $w_{ij} = 1$ otherwise, for $i = 1,2,\ldots,n$; $j = 1,2,\ldots,m$.

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_{ij} x_{ij}}{\sum_{i=1}^{n} w_{ij}}, \qquad j = 1,2,\ldots,m.$$

(b) Standard deviations:

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

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

$$\tilde{S}_{jk} = \sum_{i=1}^{n} w_{ij} w_{ik} x_{ij} x_{ik}, \qquad j,k = 1,2,\ldots,m.$$

(d) Correlation-like coefficients:

$$\tilde{R}_{jk} = \frac{\tilde{S}_{jk}}{\sqrt{\tilde{S}_{jj(k)} j(k) \tilde{S}_{kk(j)}}}, \qquad j,k = 1,2,\ldots,m,$$

where $\tilde{S}_{jj(k)} = \sum_{i=1}^{n} w_{ij} w_{ik} x_{ij}^2$ and $\tilde{S}_{kk(j)} = \sum_{i=1}^{n} w_{ik} w_{ij} x_{ik}^2$

(i.e., the sums of squares about zero are based on the same set of observations as are used in the calculation of the numerator).

If $\tilde{S}_{jj(k)}$ or $\tilde{S}_{kk(j)}$ is zero, $\tilde{R}_{jk}$ is set to zero.

(e) The number of cases used in the calculation of each of the correlation-like coefficients:

$$c_{jk} = \sum_{i=1}^{n} w_{ij} w_{ik}, \qquad j,k = 1,2,\ldots,m.$$

(The diagonal terms, c_{jj}, for $j = 1,2,\ldots,m$, also give the number of cases used in the calculation of the means $\bar{x}_j$ and the standard deviations s_j.)

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**.

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**

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**

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).

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

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldsspz, ldrz, ldcnt

5.4 Output Parameters

1: **xbar(m) – double array**

The mean value, $\bar{x}_j$, of the jth variable, for $j = 1, 2, \ldots, m$.

2: **std(m) – double array**

The standard deviation, s_j, of the jth variable, for $j = 1, 2, \ldots, m$.

3: **sspz(ldsspz,m) – double array**

sspz(j, k) is the cross-product about zero, $\tilde{S}_{jk}$, for $j, k = 1, 2, \ldots, m$.

4: **rz(ldrz,m) – double array**

rz(j, k) is the correlation-like coefficient, $\tilde{R}_{jk}$, between the jth and kth variables, for $j, k = 1, 2, \ldots, m$.

5: **ncases – int32 scalar**

The minimum number of cases used in the calculation of any of the sums of squares and cross-products and correlation-like coefficients (when cases involving missing values have been eliminated).

6: **cnt(ldcnt,m) – double array**

cnt(j, k) is the number of cases, c_{jk}, actually used in the calculation of $\tilde{S}_{jk}$, and $\tilde{R}_{jk}$, the sum of cross-products and correlation-like coefficient for the jth and kth variables, for $j, k = 1, 2, \ldots, m$.

7: **ifail – int32 scalar**

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

6 Error Indicators and Warnings

Note: g02bf may return useful information for one or more of the following detected errors or warnings.

ifail $= 1$

On entry, **n** < 2.

ifail $= 2$

On entry, **m** < 2.

ifail $= 3$

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

ifail $= 4$

After observations with missing values were omitted, fewer than two cases remained for at least one pair of variables. (The pairs of variables involved can be determined by examination of the contents of the array **cnt**). All means, standard deviations, sums of squares and cross-products, and correlation-like coefficients based on two or more cases are returned by the function even if **ifail** $= 4$.

7 Accuracy

g02bf 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. g02bf treats all values in the inclusive range $(1 \pm \mathrm{ACC}) \times xm_j$, 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 g02bf 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(1);
     int32(1)];
xmiss = [0;
     -1;
     0];
[xbar, std, sspz, rz, ncases, count, ifail] = g02bf(n, x, miss, xmiss)
```

```
xbar =
    6.7500
    7.5000
    3.5000
std =
    4.5735
    3.8730
    1.2910
sspz =
   245   111    82
   111   270    57
    82    57    54
rz =
    1.0000    0.9840    0.9055
    0.9840    1.0000    0.7699
    0.9055    0.7699    1.0000
ncases =
           3
count =
     4     3     3
     3     4     3
     3     3     4
ifail =
           0
```