NAG Library Routine Document

G07BBF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G07BBF computes maximum likelihood estimates and their standard errors for parameters of the Normal distribution from grouped and/or censored data.

2 Specification

```
SUBROUTINE G07BBF (METHOD, N, X, XC, IC, XMU, XSIG, TOL, MAXIT, SEXMU,          &
                   SEXSIG, CORR, DEV, NOBS, NIT, WK, IFAIL)

INTEGER            N, IC(N), MAXIT, NOBS(4), NIT, IFAIL
REAL (KIND=nag_wp) X(N), XC(N), XMU, XSIG, TOL, SEXMU, SEXSIG, CORR, DEV,      &
                   WK(2*N)
CHARACTER(1)       METHOD
```

3 Description

A sample of size n is taken from a Normal distribution with mean μ and variance σ^2 and consists of grouped and/or censored data. Each of the n observations is known by a pair of values (L_i, U_i) such that:

$$L_i \leq x_i \leq U_i.$$

The data is represented as particular cases of this form:

exactly specified observations occur when $L_i = U_i = x_i$,

right-censored observations, known only by a lower bound, occur when $U_i \rightarrow \infty$,

left-censored observations, known only by a upper bound, occur when $L_i \rightarrow -\infty$,

and interval-censored observations when $L_i < x_i < U_i$.

Let the set A identify the exactly specified observations, sets B and C identify the observations censored on the right and left respectively, and set D identify the observations confined between two finite limits. Also let there be r exactly specified observations, i.e., the number in A. The probability density function for the standard Normal distribution is

$$Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right), \qquad -\infty < x < \infty$$

and the cumulative distribution function is

$$P(X) = 1 - Q(X) = \int_{-\infty}^{X} Z(x)\,dx.$$

The log-likelihood of the sample can be written as:

$$L(\mu, \sigma) = -r \log \sigma - \tfrac{1}{2}\sum_{A} \{(x_i - \mu)/\sigma\}^2 + \sum_{B} \log(Q(l_i)) + \sum_{C} \log(P(u_i)) + \sum_{D} \log(p_i)$$

where $p_i = P(u_i) - P(l_i)$ and $u_i = (U_i - \mu)/\sigma, \qquad l_i = (L_i - \mu)/\sigma.$

Let

$$S(x_i) = \frac{Z(x_i)}{Q(x_i)}, \qquad S_1(l_i, u_i) = \frac{Z(l_i) - Z(u_i)}{p_i}$$

and

$$S_2(l_i, u_i) = \frac{u_i Z(u_i) - l_i Z(l_i)}{p_i},$$

then the first derivatives of the log-likelihood can be written as:

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = L_1(\mu,\sigma) = \sigma^{-2}\sum_A (x_i - \mu) + \sigma^{-1}\sum_B S(l_i) - \sigma^{-1}\sum_C S(-u_i) + \sigma^{-1}\sum_D S_1(l_i, u_i)$$

and

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = L_2(\mu,\sigma) = -r\sigma^{-1} + \sigma^{-3}\sum_A (x_i - \mu)^2 + \sigma^{-1}\sum_B l_i S(l_i) - \sigma^{-1}\sum_C u_i S(-u_i)$$

$$-\sigma^{-1}\sum_D S_2(l_i, u_i)$$

The maximum likelihood estimates, $\hat{\mu}$ and $\hat{\sigma}$, are the solution to the equations:

$$L_1(\hat{\mu}, \hat{\sigma}) = 0 \tag{1}$$

and

$$L_2(\hat{\mu}, \hat{\sigma}) = 0 \tag{2}$$

and if the second derivatives $\frac{\partial^2 L}{\partial^2 \mu}$, $\frac{\partial^2 L}{\partial \mu \partial \sigma}$ and $\frac{\partial^2 L}{\partial^2 \sigma}$ are denoted by L_{11}, L_{12} and L_{22} respectively, then estimates of the standard errors of $\hat{\mu}$ and $\hat{\sigma}$ are given by:

$$\text{se}(\hat{\mu}) = \sqrt{\frac{-L_{22}}{L_{11}L_{22} - L_{12}^2}}, \qquad \text{se}(\hat{\sigma}) = \sqrt{\frac{-L_{11}}{L_{11}L_{22} - L_{12}^2}}$$

and an estimate of the correlation of $\hat{\mu}$ and $\hat{\sigma}$ is given by:

$$\frac{L_{12}}{\sqrt{L_{12}L_{22}}}.$$

To obtain the maximum likelihood estimates the equations (1) and (2) can be solved using either the Newton–Raphson method or the Expectation-maximization (EM) algorithm of Dempster *et al.* (1977).

Newton–Raphson Method

This consists of using approximate estimates $\tilde{\mu}$ and $\tilde{\sigma}$ to obtain improved estimates $\tilde{\mu} + \delta\tilde{\mu}$ and $\tilde{\sigma} + \delta\tilde{\sigma}$ by solving

$$\delta\tilde{\mu}L_{11} + \delta\tilde{\sigma}L_{12} + L_1 = 0,$$

$$\delta\tilde{\mu}L_{12} + \delta\tilde{\sigma}L_{22} + L_2 = 0,$$

for the corrections $\delta\tilde{\mu}$ and $\delta\tilde{\sigma}$.

EM Algorithm

The expectation step consists of constructing the variable w_i as follows:

$$\text{if} \quad i \in A, \quad w_i = x_i \tag{3}$$

$$\text{if} \quad i \in B, \quad w_i = E(x_i \mid x_i > L_i) = \mu + \sigma S(l_i) \tag{4}$$

$$\text{if} \quad i \in C, \quad w_i = E(x_i \mid x_i < U_i) = \mu - \sigma S(-u_i) \tag{5}$$

$$\text{if} \quad i \in D, \quad w_i = E(x_i \mid L_i < x_i < U_i) = \mu + \sigma S_1(l_i, u_i) \tag{6}$$

the maximization step consists of substituting (3), (4), (5) and (6) into (1) and (2) giving:

$$\hat{\mu} = \sum_{i=1}^{n} \hat{w}_i / n \tag{7}$$

and

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (\hat{w}_i - \hat{\mu})^2 / \left\{ r + \sum_{B} T\left(\hat{l}_i\right) + \sum_{C} T(-\hat{u}_i) + \sum_{D} T_1\left(\hat{l}_i, \hat{u}_i\right) \right\} \tag{8}$$

where

$$T(x) = S(x)\{S(x) - x\}, \qquad T_1(l,u) = S_1^2(l,u) + S_2(l,u)$$

and where $\hat{w}_i$, $\hat{l}_i$ and $\hat{u}_i$ are w_i, l_i and u_i evaluated at $\hat{\mu}$ and $\hat{\sigma}$. Equations (3) to (8) are the basis of the EM iterative procedure for finding $\hat{\mu}$ and $\hat{\sigma}^2$. The procedure consists of alternately estimating $\hat{\mu}$ and $\hat{\sigma}^2$ using (7) and (8) and estimating $\{\hat{w}_i\}$ using (3) to (6).

In choosing between the two methods a general rule is that the Newton–Raphson method converges more quickly but requires good initial estimates whereas the EM algorithm converges slowly but is robust to the initial values. In the case of the censored Normal distribution, if only a small proportion of the observations are censored then estimates based on the exact observations should give good enough initial estimates for the Newton–Raphson method to be used. If there are a high proportion of censored observations then the EM algorithm should be used and if high accuracy is required the subsequent use of the Newton–Raphson method to refine the estimates obtained from the EM algorithm should be considered.

4 References

Dempster A P, Laird N M and Rubin D B (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion) *J. Roy. Statist. Soc. Ser. B* **39** 1–38

Swan A V (1969) Algorithm AS 16. Maximum likelihood estimation from grouped and censored normal data *Appl. Statist.* **18** 110–114

Wolynetz M S (1979) Maximum likelihood estimation from confined and censored normal data *Appl. Statist.* **28** 185–195

5 Parameters

1: METHOD – CHARACTER(1) *Input*

On entry: indicates whether the Newton–Raphson or EM algorithm should be used.

If METHOD = 'N', then the Newton–Raphson algorithm is used.

If METHOD = 'E', then the EM algorithm is used.

Constraint: METHOD = 'N' or 'E'.

2: N – INTEGER *Input*

On entry: n, the number of observations.

Constraint: N $\geq$ 2.

3: X(N) – REAL (KIND=nag_wp) array *Input*

On entry: the observations x_i, L_i or U_i, for $i = 1, 2, \ldots, n$.

If the observation is exactly specified – the exact value, x_i.

If the observation is right-censored – the lower value, L_i.

If the observation is left-censored – the upper value, U_i.

If the observation is interval-censored – the lower or upper value, L_i or U_i, (see XC).

4: XC(N) – REAL (KIND=nag_wp) array *Input*

On entry: if the jth observation, for $j = 1, 2, \ldots, n$ is an interval-censored observation then XC(j) should contain the complementary value to X(j), that is, if X(j) < XC(j), then XC(j) contains upper value, U_i, and if X(j) > XC(j), then XC(j) contains lower value, L_i. Otherwise if the jth observation is exact or right- or left-censored XC(j) need not be set.

Note: if X(j) = XC(j) then the observation is ignored.

5: IC(N) – INTEGER array *Input*

On entry: IC(i) contains the censoring codes for the ith observation, for $i = 1, 2, \ldots, n$.

If IC(i) = 0, the observation is exactly specified.

If IC(i) = 1, the observation is right-censored.

If IC(i) = 2, the observation is left-censored.

If IC(i) = 3, the observation is interval-censored.

Constraint: IC(i) = 0, 1, 2 or 3, for $i = 1, 2, \ldots, n$.

6: XMU – REAL (KIND=nag_wp) *Input/Output*

On entry: if XSIG > 0.0 the initial estimate of the mean, μ; otherwise XMU need not be set.

On exit: the maximum likelihood estimate, $\hat{\mu}$, of μ.

7: XSIG – REAL (KIND=nag_wp) *Input/Output*

On entry: specifies whether an initial estimate of μ and σ are to be supplied.

XSIG > 0.0
XSIG is the initial estimate of σ and XMU must contain an initial estimate of μ.

XSIG ≤ 0.0
Initial estimates of XMU and XSIG are calculated internally from:

(a) the exact observations, if the number of exactly specified observations is ≥ 2; or

(b) the interval-censored observations; if the number of interval-censored observations is ≥ 1; or

(c) they are set to 0.0 and 1.0 respectively.

On exit: the maximum likelihood estimate, $\hat{\sigma}$, of σ.

8: TOL – REAL (KIND=nag_wp) *Input*

On entry: the relative precision required for the final estimates of μ and σ. Convergence is assumed when the absolute relative changes in the estimates of both μ and σ are less than TOL.

If TOL = 0.0, then a relative precision of 0.000005 is used.

Constraint: ***machine precision*** < TOL ≤ 1.0 or TOL = 0.0.

9: MAXIT – INTEGER *Input*

On entry: the maximum number of iterations.

If MAXIT ≤ 0, then a value of 25 is used.

10: SEXMU – REAL (KIND=nag_wp) *Output*

On exit: the estimate of the standard error of $\hat{\mu}$.

11: SEXSIG – REAL (KIND=nag_wp) *Output*

On exit: the estimate of the standard error of $\hat{\sigma}$.

12: CORR – REAL (KIND=nag_wp) *Output*

On exit: the estimate of the correlation between $\hat{\mu}$ and $\hat{\sigma}$.

13: DEV – REAL (KIND=nag_wp) *Output*

On exit: the maximized log-likelihood, $L(\hat{\mu}, \hat{\sigma})$.

14: NOBS(4) – INTEGER array *Output*

On exit: the number of the different types of each observation;

NOBS(1) contains number of right-censored observations.

NOBS(2) contains number of left-censored observations.

NOBS(3) contains number of interval-censored observations.

NOBS(4) contains number of exactly specified observations.

15: NIT – INTEGER *Output*

On exit: the number of iterations performed.

16: WK($2 \times$ N) – REAL (KIND=nag_wp) array *Workspace*

17: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

6 Error Indicators and Warnings

If on entry IFAIL $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, METHOD $\neq$ 'N' or 'E',
or N < 2,
or IC(i) $\neq 0$, 1, 2 or 3, for some i,
or TOL < 0.0,
or $0.0 <$ TOL $<$ ***machine precision***,
or TOL > 1.0.

IFAIL $= 2$

The chosen method failed to converge in MAXIT iterations. You should either increase TOL or MAXIT or, if using the EM algorithm try using the Newton–Raphson method with initial values those returned by the current call to G07BBF. All returned values will be reasonable approximations to the correct results if MAXIT is not very small.

IFAIL $= 3$

The chosen method is diverging. This will be due to poor initial values. You should try different initial values.

IFAIL $= 4$

G07BBF was unable to calculate the standard errors. This can be caused by the method starting to diverge when the maximum number of iterations was reached.

7 Accuracy

The accuracy is controlled by the parameter TOL.

If high precision is requested with the EM algorithm then there is a possibility that, due to the slow convergence, before the correct solution has been reached the increments of $\hat{\mu}$ and $\hat{\sigma}$ may be smaller than TOL and the process will prematurely assume convergence.

8 Further Comments

The process is deemed divergent if three successive increments of μ or σ increase.

9 Example

A sample of 18 observations and their censoring codes are read in and the Newton–Raphson method used to compute the estimates.

9.1 Program Text

```
    Program g07bbfe

!     G07BBF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g07bbf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: corr, dev, sexmu, sexsig, tol, xmu,  &
                                          xsig
      Integer                          :: i, ifail, maxit, n, nit
      Character (1)                    :: method
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: wk(:), x(:), xc(:)
      Integer, Allocatable             :: ic(:)
      Integer                          :: nobs(4)
!     .. Executable Statements ..
      Write (nout,*) 'G07BBF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in problem size and control parameters
      Read (nin,*) n, method, xmu, xsig, tol, maxit

      Allocate (x(n),xc(n),ic(n),wk(2*n))

!     Read in data
      Read (nin,*)(x(i),xc(i),ic(i),i=1,n)

!     Calculate estimates
```

```
      ifail = 0
      Call g07bbf(method,n,x,xc,ic,xmu,xsig,tol,maxit,sexmu,sexsig,corr,dev, &
        nobs,nit,wk,ifail)

!     Display results
      Write (nout,99999) ' Mean = ', xmu
      Write (nout,99999) ' Standard deviation = ', xsig
      Write (nout,99999) ' Standard error of mean = ', sexmu
      Write (nout,99999) ' Standard error of sigma = ', sexsig
      Write (nout,99999) ' Correlation coefficient = ', corr
      Write (nout,99998) ' Number of right censored observations = ', nobs(1)
      Write (nout,99998) ' Number of left censored observations = ', nobs(2)
      Write (nout,99998) ' Number of interval censored observations = ', &
        nobs(3)
      Write (nout,99998) ' Number of exactly specified observations = ', &
        nobs(4)
      Write (nout,99998) ' Number of iterations = ', nit
      Write (nout,99999) ' Log-likelihood = ', dev

99999 Format (1X,A,F8.4)
99998 Format (1X,A,I2)
    End Program g07bbfe
```

9.2 Program Data

```
G07BBF Example Program Data
18 'N' 4.0 1.0 0.00005 50
4.5 0.0 0 5.4 0.0 0 3.9 0.0 0 5.1 0.0 0 4.6 0.0 0 4.8 0.0 0
2.9 0.0 0 6.3 0.0 0 5.5 0.0 0 4.6 0.0 0 4.1 0.0 0 5.2 0.0 0
3.2 0.0 1 4.0 0.0 1 3.1 0.0 1 5.1 0.0 2 3.8 0.0 2 2.2 2.5 3
```

9.3 Program Results

```
 G07BBF Example Program Results

  Mean =   4.4924
  Standard deviation =   1.0196
  Standard error of mean =   0.2606
  Standard error of sigma =   0.1940
  Correlation coefficient =   0.0160
  Number of right censored observations =  3
  Number of left censored observations =  2
  Number of interval censored observations =  1
  Number of exactly specified observations = 12
  Number of iterations =  5
  Log-likelihood = -22.2817
```