

## NAG Toolbox

### nag\_univar\_estim\_normal (g07bb)

#### 1 Purpose

nag\_univar\_estim\_normal (g07bb) computes maximum likelihood estimates and their standard errors for arguments of the Normal distribution from grouped and/or censored data.

#### 2 Syntax

```
[xmu, xsig, sexmu, sexsig, corr, dev, nob, nit, ifail] =
nag_univar_estim_normal(method, x, xc, ic, xmu, xsig, tol, maxit, 'n', n)

[xmu, xsig, sexmu, sexsig, corr, dev, nob, nit, ifail] = g07bb(method, x, xc,
ic, xmu, xsig, tol, maxit, 'n', n)
```

#### 3 Description

A sample of size  $n$  is taken from a Normal distribution with mean  $\mu$  and variance  $\sigma^2$  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 an 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

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

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:

$$L(\mu, \sigma) = -r \log \sigma - \frac{1}{2} \sum_A \{(x_i - \mu)/\sigma\}^2 + \sum_B \log(Q(l_i)) + \sum_C \log(P(u_i)) + \sum_D \log(p_i)$$

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

Let

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

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

$$\begin{aligned} \frac{\partial L(\mu, \sigma)}{\partial \sigma} = L_2(\mu, \sigma) = & -r\sigma^{-1} + \sigma^{-3} \sum_A (x_i - \mu)^2 + \sigma^{-1} \sum_B l_i S(l_i) - \sigma^{-1} \sum_C u_i S(-u_i) \\ & - \sigma^{-1} \sum_D S_2(l_i, u_i) \end{aligned}$$

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

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

and

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

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:

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

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

$$\frac{L_{12}}{\sqrt{L_{11}L_{22}}}$$

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:

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

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

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

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

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

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

and

$$\hat{\sigma}^2 = \sum_{i=1}^n (\hat{w}_i - \hat{\mu})^2 / \left\{ r + \sum_B T(\hat{l}_i) + \sum_C T(-\hat{u}_i) + \sum_D T_1(\hat{l}_i, \hat{u}_i) \right\} \quad (8)$$

where

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

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

### 5.1 Compulsory Input Parameters

1: **method** – CHARACTER(1)

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: **x(n)** – REAL (KIND=nag\_wp) array

The observations  $x_i$ ,  $L_i$  or  $U_i$ , for  $i = 1, 2, \dots, n$ .

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**).

3: **xc(n)** – REAL (KIND=nag\_wp) array

If the  $j$ th observation, for  $j = 1, 2, \dots, n$  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  $j$ th observation is exact or right- or left-censored **xc(j)** need not be set.

**Note:** if  $\mathbf{x}(j) = \mathbf{xc}(j)$  then the observation is ignored.

4: **ic(n)** – INTEGER array

**ic(i)** contains the censoring codes for the  $i$ th observation, for  $i = 1, 2, \dots, n$ .

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  $i = 1, 2, \dots, n$ .

5: **xmu** – REAL (KIND=nag\_wp)

If **xsig** > 0.0 the initial estimate of the mean,  $\mu$ ; otherwise **xmu** need not be set.

6: **xsig** – REAL (KIND=nag\_wp)

Specifies whether an initial estimate of  $\mu$  and  $\sigma$  are to be supplied.

**xsig** > 0.0

**xsig** is the initial estimate of  $\sigma$  and **xmu** must contain an initial estimate of  $\mu$ .

**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  $\geq 2$ ; or

(b) the interval-censored observations; if the number of interval-censored observations is  $\geq 1$ ; or

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

7: **tol** – REAL (KIND=nag\_wp)

The relative precision required for the final estimates of  $\mu$  and  $\sigma$ . Convergence is assumed when the absolute relative changes in the estimates of both  $\mu$  and  $\sigma$  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.

8: **maxit** – INTEGER

The maximum number of iterations.

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

## 5.2 Optional Input Parameters

1: **n** – INTEGER

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

$n$ , the number of observations.

*Constraint:* **n** ≥ 2.

## 5.3 Output Parameters

1: **xmu** – REAL (KIND=nag\_wp)

The maximum likelihood estimate,  $\hat{\mu}$ , of  $\mu$ .

- 2: **xsig** – REAL (KIND=nag\_wp)  
The maximum likelihood estimate,  $\hat{\sigma}$ , of  $\sigma$ .
- 3: **sexmu** – REAL (KIND=nag\_wp)  
The estimate of the standard error of  $\hat{\mu}$ .
- 4: **sexsig** – REAL (KIND=nag\_wp)  
The estimate of the standard error of  $\hat{\sigma}$ .
- 5: **corr** – REAL (KIND=nag\_wp)  
The estimate of the correlation between  $\hat{\mu}$  and  $\hat{\sigma}$ .
- 6: **dev** – REAL (KIND=nag\_wp)  
The maximized log-likelihood,  $L(\hat{\mu}, \hat{\sigma})$ .
- 7: **nobs(4)** – INTEGER array  
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.
- 8: **nit** – INTEGER  
The number of iterations performed.
- 9: **ifail** – INTEGER  
**ifail** = 0 unless the function detects an error (see Section 5).

## 6 Error Indicators and Warnings

Errors or warnings detected by the function:

**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  $0.0 < \mathbf{tol} < \mathit{machine\ precision}$ ,  
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 nag\_univar\_estim\_normal (g07bb). 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

nag\_univar\_estim\_normal (g07bb) 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.

**ifail** = -99

An unexpected error has been triggered by this routine. Please contact NAG.

**ifail** = -399

Your licence key may have expired or may not have been installed correctly.

**ifail** = -999

Dynamic memory allocation failed.

## 7 Accuracy

The accuracy is controlled by the argument **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  $\mu$  or  $\sigma$  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

```
function g07bb_example

fprintf('g07bb example results\n\n');

% Data
x = [4.5;    5.4;    3.9;    5.1;    4.6;    4.8;
     2.9;    6.3;    5.5;    4.6;    4.1;    5.2;
     3.2;    4;     3.1;    5.1;    3.8;    2.2];

n = numel(x);
xc = zeros(n,1);
ic = zeros(n,1,nag_int_name);
xc(n,1) = 2.5;
ic(n-5:n,1) = [1; 1; 1; 2; 2; 3];

% Parameters
method = 'N';
xmu = 4;
xsig = 1;
tol = 5e-05;
maxit = nag_int(50);

% Calculate estimates
[xmu, xsig, sexmu, sexsig, corr, dev, nobs, nit, ifail] = ...
g07bb( ...
    method, x, xc, ic, xmu, xsig, tol, maxit);

% Display results
fprintf(' Mean                = %8.4f\n', xmu);
fprintf(' Standard deviation    = %8.4f\n', xsig);
```

```

fprintf(' Standard error of mean           = %8.4f\n', sexmu);
fprintf(' Standard error of sigma         = %8.4f\n', sexsig);
fprintf(' Correlation coefficient             = %8.4f\n', corr);
fprintf(' Number of right censored observations = %3d\n', nobs(1));
fprintf(' Number of left censored observations  = %3d\n', nobs(2));
fprintf(' Number of interval censored observations = %3d\n', nobs(3));
fprintf(' Number of exactly specified observations = %3d\n', nobs(4));
fprintf(' Number of iterations                 = %3d\n', nit);
fprintf(' Log-likelihood                       = %8.4f\n', dev);

```

## 9.2 Program Results

g07bb example 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

```

---