

NAG C Library Function Document

nag_censored_normal (g07bbc)

1 Purpose

nag_censored_normal (g07bbc) computes maximum likelihood estimates and their standard errors for parameters of the Normal distribution from grouped and/or censored data.

2 Specification

```
void nag_censored_normal (Nag_CEMethod method, Integer n, const double x[],
    const double xc[], const Integer ic[], double *xmu, double *xsig, double tol,
    Integer maxit, double *sexmu, double *sexsig, double *corr, double *dev,
    Integer nobs[], Integer *nit, NagError *fail)
```

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 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 AS16. 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** – Nag_CEMethod *Input*
On entry: indicates whether the Newton–Raphson or *EM* algorithm should be used.
 If **method** = **Nag-CE-NR**, then the Newton–Raphson algorithm is used.
 If **method** = **Nag-CE-EM**, then the *EM* algorithm is used.
Constraint: **method** = **Nag-CE-NR** or **Nag-CE-EM**.
- 2: **n** – Integer *Input*
On entry: the number of observations, n .
Constraint: $n \geq 2$.
- 3: **x[n]** – const double *Input*
On entry: 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**).
- 4: **xc[n]** – const double *Input*
On entry: if the j th observation, for $j = 1, 2, \dots, n$ is an interval-censored observation then **xc[j – 1]** should contain the complementary value to **x[j – 1]**, that is, if **x[j – 1] < xc[j – 1]**, then

$\mathbf{xc}[j-1]$ contains upper value, U_i , and if $\mathbf{x}[j-1] > \mathbf{xc}[j-1]$, then $\mathbf{xc}[j-1]$ contains lower value, L_i . Otherwise if the j th observation is exact or right- or left-censored $\mathbf{xc}[j-1]$ need not be set.

Note: if $\mathbf{x}[j-1] = \mathbf{xc}[j-1]$ then the observation is ignored.

- 5: **ic[n]** – const Integer *Input*
On entry: **ic**[$i-1$] contains the censoring codes for the i th observation, for $i = 1, 2, \dots, n$.
 If **ic**[$i-1$] = 0, the observation is exactly specified.
 If **ic**[$i-1$] = 1, the observation is right-censored.
 If **ic**[$i-1$] = 2, the observation is left-censored.
 If **ic**[$i-1$] = 3, the observation is interval-censored.
Constraint: **ic**[$i-1$] = 0, 1, 2 or 3, for $i = 1, 2, \dots, n$.
- 6: **xmu** – double * *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** – double * *Input/Output*
On entry: specifies whether an initial estimate of μ and σ are to be supplied. If **xsig** > 0.0, then **xsig** is the initial estimate of σ and **xmu** must contain an initial estimate of μ .
 If **xsig** ≤ 0.0, then 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** – double *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** – double * *Output*
On exit: the estimate of the standard error of $\hat{\mu}$.
- 11: **sexsig** – double * *Output*
On exit: the estimate of the standard error of $\hat{\sigma}$.
- 12: **corr** – double * *Output*
On exit: the estimate of the correlation between $\hat{\mu}$ and $\hat{\sigma}$.
- 13: **dev** – double * *Output*
On exit: the maximized log-likelihood, $L(\hat{\mu}, \hat{\sigma})$.

- 14: **nobs**[4] – Integer *Output*
On exit: the number of the different types of each observation;
nobs[0] contains number of right-censored observations.
nobs[1] contains number of left-censored observations.
nobs[2] contains number of interval-censored observations.
nobs[3] contains number of exactly specified observations.
- 15: **nit** – Integer * *Output*
On exit: the number of iterations performed.
- 16: **fail** – NagError * *Input/Output*
The NAG error parameter (see the Essential Introduction).

6 Error Indicators and Warnings

NE_INT

On entry, **n** = $\langle value \rangle$.
Constraint: **n** \geq 2.

NE_CONVERGENCE

Method has not converged in $\langle value \rangle$ iterations.

NE_DIVERGENCE

Process has diverged.

NE_EM_PROCESS

The EM process has failed.

NE_OBSERVATIONS

On entry, effective number of observations < 2 .

NE_REAL

On entry, **tol** is invalid: **tol** = $\langle value \rangle$.

NE_STANDARD_ERRORS

Standard errors cannot be computed.

NE_ALLOC_FAIL

Memory allocation failed.

NE_BAD_PARAM

On entry, parameter $\langle value \rangle$ had an illegal value.

NE_INTERNAL_ERROR

An internal error has occurred in this function. Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

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

```

/* nag_censored_normal (g07bbc) Example Program.
 *
 * Copyright 2001 Numerical Algorithms Group.
 *
 * Mark 7, 2001.
 */

#include <stdio.h>
#include <string.h>
#include <nag.h>
#include <nag_stdlib.h>
#include <nagg07.h>

int main(void)
{
    /* Scalars */
    double corr, dev, sexmu, sexsig, tol, xmu, xsig;
    Integer exit_status, i, maxit, n, nit;

    /* Arrays */
    char *method=0;
    double *x=0, *xc=0;
    Integer *ic=0, *nobs=0;
    NagError fail;
    Nag_CEMethod method_enum;

    INIT_FAIL(fail);
    exit_status = 0;
    Vprintf("g07bbc Example Program Results\n");

    /* Skip heading in data file */
    Vscanf("%*[^\\n] ");

    /* Allocate memory */
    if ( !(method = NAG_ALLOC(2, char)) )
    {
        Vprintf("Allocation failure\n");
        exit_status = -1;
        goto END;
    }

    Vscanf("%ld ' %1s '%lf%lf%lf%ld%*[^\\n] ", &n, method, &xmu, &xsig, &tol, &max-
it);

    /* Allocate memory */
    if ( !(x = NAG_ALLOC(n, double)) ||
        !(xc = NAG_ALLOC(n, double)) ||

```

```

        !(ic = NAG_ALLOC(n, Integer)) ||
        !(nobs = NAG_ALLOC(4, Integer)) )
    {
        Vprintf("Allocation failure\n");
        exit_status = -1;
        goto END;
    }

    for (i = 1; i <= n; ++i)
        Vscanf("%lf%lf%ld", &x[i - 1], &xc[i - 1], &ic[i - 1]);
    Vscanf("%*[\n] ");

    if (!(strcmp(method, "N")))
        method_enum = Nag_CE_NR;
    else if (!(strcmp(method, "E")))
        method_enum = Nag_CE_EM;
    else
    {
        Vprintf("Invalid method\n");
        exit_status = -1;
        goto END;
    }
    g07bbc(method_enum, n, x, xc, ic, &xmu, &xsig, tol, maxit, &sexmu,
        &sexsig, &corr, &dev, nobs, &nit, &fail);

    if (fail.code != NE_NOERROR)
    {
        Vprintf("Error from g07bbc.\n%s\n", fail.message);
        exit_status = 1;
        goto END;
    }

    Vprintf("\n");
    Vprintf(" Mean = %8.4f\n", xmu);
    Vprintf(" Standard deviation = %8.4f\n", xsig);
    Vprintf(" Standard error of mean = %8.4f\n", sexmu);
    Vprintf(" Standard error of sigma = %8.4f\n", sexsig);
    Vprintf(" Correlation coefficient = %8.4f\n", corr);
    Vprintf(" Number of right censored observations = %2ld\n", nobs[0]);
    Vprintf(" Number of left censored observations = %2ld\n", nobs[1]);
    Vprintf(" Number of interval censored observations = %2ld\n", nobs[2]);
    Vprintf(" Number of exactly specified observations = %2ld\n", nobs[3]);
    Vprintf(" Number of iterations = %2ld\n", nit);
    Vprintf(" Log-likelihood = %8.4f\n", dev);

    END:
    if (method) NAG_FREE(method);
    if (x) NAG_FREE(x);
    if (xc) NAG_FREE(xc);
    if (ic) NAG_FREE(ic);
    if (nobs) NAG_FREE(nobs);

    return exit_status;
}

```

9.2 Program Data

g07bbc 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

g07bbc 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
