# G07EBF – NAG Fortran Library Routine Document

**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

G07EBF calculates a rank based (nonparametric) estimate and confidence interval for the difference in location between two independent populations.

# 2 Specification

```
      SUBROUTINE G07EBF(METHOD, N, X, M, Y, CLEVEL, THETA, THETAL,
     1                  THETAU, ESTCL, ULOWER, UUPPER, WRK, IWRK, IFAIL)
      INTEGER           N, M, IWRK(3*N), IFAIL
      real              X(N), Y(M), CLEVEL, THETA, THETAL, THETAU,
     1                  ESTCL, ULOWER, UUPPER, WRK(3*(M+N))
      CHARACTER*1       METHOD
```

# 3 Description

Consider two random samples from two populations which have the same continuous distribution except for a shift in the location. Let the random sample, $x = (x_1, x_2, \ldots, x_n)^T$, have distribution $F(x)$ and the random sample, $y = (y_1, y_2, \ldots, y_m)^T$, have distribution $F(x - \theta)$.

G07EBF finds a point estimate, $\hat{\theta}$, of the difference in location $\theta$ together with an associated confidence interval. The estimates are based on the ordered differences $y_j - x_i$. The estimate $\hat{\theta}$ is defined by

$$\hat{\theta} = \text{median}\{y_j - x_i, i = 1, 2, \ldots, n; j = 1, 2, \ldots, m\}.$$

Let $d_k$ for $k = 1, 2, \ldots, nm$ denote the $nm$ (ascendingly) ordered differences $y_j - x_i$ for $i = 1, 2, \ldots, n; j = 1, 2, \ldots, m$. Then

if $nm$ is odd, $\hat{\theta} = d_k$ where $k = (nm - 1)/2$,

if $nm$ is even, $\hat{\theta} = (d_k + d_{k+1})/2$ where $k = nm/2$.

This estimator arises from inverting the two sample Mann–Whitney rank test statistic, $U(\theta_0)$, for testing the hypothesis that $\theta = \theta_0$. Thus $U(\theta_0)$ is the value of the Mann–Whitney $U$ statistic for the two independent samples $\{(x_i + \theta_0), i = 1, 2 \ldots, n\}$ and $\{y_j, j = 1, 2 \ldots, m\}$. Effectively $U(\theta_0)$ is a monotonically increasing step function of $\theta_0$ with

$$\begin{aligned} \text{mean}(U) &= \mu = \frac{nm}{2}, \\ \text{var}(U) &= \sigma^2 = \frac{nm(n + m + 1)}{12}. \end{aligned}$$

The estimate $\hat{\theta}$ is the solution to the equation $U(\hat{\theta}) = \mu$; two methods are available for solving this equation. These methods avoid the computation of all the ordered differences $d_k$; this is because for large $n$ and $m$ both the storage requirements and the computation time would be high.

The first is an exact method based on a set partitioning procedure on the set of all differences $y_j - x_i$ for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$. This is adapted from the algorithm proposed by Monahan [3] for the computation of the Hodges–Lehmann estimator for a single population.

The second is an iterative algorithm, based on the Illinois method which is a modification of the regula falsi method, see McKean and Ryan [2]. This algorithm has proved suitable for the function $U(\theta_0)$ which is asymptotically linear as a function of $\theta_0$.

The confidence interval limits are also based on the inversion of the Mann–Whitney test statistic.

Given a desired percentage for the confidence interval, $1 - \alpha$, expressed as a proportion between 0.0 and 1.0 initial estimates of the upper and lower confidence limits for the Mann–Whitney $U$ statistic are found;

$$\begin{aligned} U_l &= \mu - 0.5 + (\sigma \times \Phi^{-1}(\alpha/2)) \\ U_u &= \mu + 0.5 + (\sigma \times \Phi^{-1}((1-\alpha)/2)) \end{aligned}$$

where $\Phi^{-1}$ is the inverse cumulative Normal distribution function.

$U_l$ and $U_u$ are rounded to the nearest integer values. These estimates are refined using an exact method, without taking ties into account, if $n + m \leq 40$ and $\max(n, m) \leq 30$ and a Normal approximation otherwise, to find $U_l$ and $U_u$ satisfying

$$\begin{aligned} &P(U \leq U_l) \leq \alpha/2 \\ &P(U \leq U_l + 1) > \alpha/2 \end{aligned}$$

and

$$\begin{aligned} &P(U \geq U_u) \leq \alpha/2 \\ &P(U \geq U_u - 1) > \alpha/2. \end{aligned}$$

The function $U(\theta_0)$ is a monotonically increasing step function. It is the number of times a score in the second sample, $y_j$, precedes a score in the first sample, $x_i + \theta$, where we only count a half if a score in the second sample actually equals a score in the first.

Let $U_l = k$; then $\theta_l = d_{k+1}$. This is the largest value $\theta_l$ such that $U(\theta_l) = U_l$.

Let $U_u = nm - k$; then $\theta_u = d_{nm-k}$. This is the smallest value $\theta_u$ such that $U(\theta_u) = U_u$.

As in the case of $\hat{\theta}$, these equations may be solved using either the exact or iterative methods to find the values $\theta_l$ and $\theta_u$.

Then $(\theta_l, \theta_u)$ is the confidence interval for $\theta$. The confidence interval is thus defined by those values of $\theta_0$ such that the null hypothesis, $\theta = \theta_0$, is not rejected by the Mann–Whitney two sample rank test at the $(100 \times \alpha)\%$ level.

# 4 References

**[1]** Lehmann E L (1975) *Nonparametrics: Statistical Methods Based on Ranks* Holden–Day

**[2]** McKean J W and Ryan T A (1977) Algorithm 516: An algorithm for obtaining confidence intervals and point estimates based on ranks in the two-sample location problem *ACM Trans. Math. Software* **10** 183–185

**[3]** Monahan J F (1984) Algorithm 616: Fast computation of the Hodges–Lehman location estimator *ACM Trans. Math. Software* **10** 265–270

# 5 Parameters

**1:** METHOD — CHARACTER*1 *Input*

*On entry:* specifies the method to be used.

If METHOD = 'E' the exact algorithm is used.

If METHOD = 'A' the iterative algorithm is used.

*Constraint:* METHOD = 'E' or 'A'.

**2:** N — INTEGER *Input*

*On entry:* the size of the first sample, $n$.

*Constraint:* N $\geq 1$.

**3:** X(N) — ***real*** array *Input*

*On entry:* the observations of the first sample, $x_i$ for $i = 1, 2, \ldots, n$.

**4:** M — INTEGER *Input*

*On entry:* the size of the second sample, $m$.

*Constraint:* M $\geq 1$.

**5:** Y(M) — ***real*** array *Input*

*On entry:* the observations of the second sample, $y_j$ for $j = 1, 2, \ldots, m$.

**6:** CLEVEL — ***real*** *Input*

*On entry:* the confidence interval required, $1 - \alpha$; e.g., for a 95% confidence interval set CLEVEL = 0.95.

*Constraint:* $0.0 <$ CLEVEL $< 1.0$.

**7:** THETA — ***real*** *Output*

*On exit:* the estimate of the difference in the location of the two populations, $\hat{\theta}$.

**8:** THETAL — ***real*** *Output*

*On exit:* the estimate of the lower limit of the confidence interval, $\theta_l$.

**9:** THETAU — ***real*** *Output*

*On exit:* the estimate of the upper limit of the confidence interval, $\theta_u$.

**10:** ESTCL — ***real*** *Output*

*On exit:* an estimate of the actual percentage confidence of the interval found, as a proportion between (0.0,1.0).

**11:** ULOWER — ***real*** *Output*

*On exit:* the value of the Mann–Whitney $U$ statistic corresponding to the lower confidence limit, $U_l$.

**12:** UUPPER — ***real*** *Output*

*On exit:* the value of the Mann–Whitney $U$ statistic corresponding to the upper confidence limit, $U_u$.

**13:** WRK(3*(M+N)) — ***real*** array *Workspace*

**14:** IWRK(3*N) — INTEGER array *Workspace*

**15:** IFAIL — INTEGER *Input/Output*

*On entry:* IFAIL must be set to 0, $-1$ or 1. For users not familiar with this parameter (described in Chapter P01) the recommended value is 0.

*On exit:* IFAIL = 0 unless the routine detects an error (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 detected by the routine:

IFAIL = 1

On entry, METHOD ≠ 'E' or 'A',
or N < 1,
or M < 1,
or CLEVEL ≤ 0.0,
or CLEVEL ≥ 1.0.

IFAIL = 2

Each sample consists of identical values. All estimates are set to the common difference between the samples.

IFAIL = 3

For at least one of the estimates $\hat{\theta}$, $\theta_l$ and $\theta_u$, the underlying iterative algorithm (when METHOD = 'A') failed to converge. This is an unlikely exit but the estimate should still be a reasonable approximation.

# 7 Accuracy

The routine should return results accurate to 5 significant figures in the width of the confidence interval, that is the error for any one of the three estimates should be less than $0.00001 \times (\mathrm{THETAU} - \mathrm{THETAL})$.

# 8 Further Comments

The time taken increases with the sample sizes $n$ and $m$.

# 9 Example

The following program calculates a 95% confidence interval for the difference in location between the two populations from which the two samples of sizes 50 and 100 are drawn respectively.

## 9.1 Program Text

**Note.** The listing of the example program presented below uses bold italicised terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G07EBF Example Program Text
*     Mark 16 Release. NAG Copyright 1992.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX
      PARAMETER        (NMAX=100,MMAX=100)
*     .. Local Scalars ..
      real             CLEVEL, ESTCL, THETA, THETAL, THETAU, ULOWER,
     +                 UUPPER
      INTEGER          I, IFAIL, M, N
*     .. Local Arrays ..
      real             WRK(3*(NMAX+MMAX)), X(NMAX), Y(MMAX)
      INTEGER          IWRK(3*NMAX)
*     .. External Subroutines ..
      EXTERNAL         G07EBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G07EBF Example Program Results'
*     Skip Heading in data file
```

```
      READ (NIN,*)
      READ (NIN,*) N, M
      IF (N.LE.1 .OR. N.GT.NMAX .OR. M.LE.1 .OR. M.GT.MMAX) THEN
         WRITE (NOUT,99999) N, M
      ELSE
         READ (NIN,*)
         READ (NIN,*) (X(I),I=1,N)
         READ (NIN,*)
         READ (NIN,*) (Y(I),I=1,M)
         READ (NIN,*)
         READ (NIN,*) CLEVEL
         IFAIL = 0
*
         CALL G07EBF('Approx',N,X,M,Y,CLEVEL,THETA,THETAL,THETAU,ESTCL,
     +               ULOWER,UUPPER,WRK,IWRK,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,*) ' Location estimator     Confidence Interval '
         WRITE (NOUT,*)
         WRITE (NOUT,99998) THETA, '( ', THETAL, ' , ', THETAU, ' )'
         WRITE (NOUT,*)
         WRITE (NOUT,*) ' Corresponding Mann-Whitney U statistics'
         WRITE (NOUT,*)
         WRITE (NOUT,99997) ' Lower : ', ULOWER
         WRITE (NOUT,99997) ' Upper : ', UUPPER
      END IF
      STOP
*
99999 FORMAT (4X,'N or M is out of range : N = ',I8,'  and M = ',I8)
99998 FORMAT (3X,F10.4,12X,A,F6.4,A,F6.4,A)
99997 FORMAT (A,F8.2)
      END
```

## 9.2 Program Data

```
G07EBF Example Program Data
 50 100
 First sample of N observations
 -0.582  0.157 -0.523 -0.769  2.338  1.664 -0.981  1.549  1.131 -0.460
 -0.484  1.932  0.306 -0.602 -0.979  0.132  0.256 -0.094  1.065 -1.084
 -0.969 -0.524  0.239  1.512 -0.782 -0.252 -1.163  1.376  1.674  0.831
  1.478 -1.486 -0.808 -0.429 -2.002  0.482 -1.584 -0.105  0.429  0.568
  0.944  2.558 -1.801  0.242  0.763 -0.461 -1.497 -1.353  0.301  1.941
 Second sample of M observations
  1.995  0.007  0.997  1.089  2.004  0.171  0.294  2.448  0.214  0.773
  2.960  0.025  0.638  0.937 -0.568 -0.711  0.931  2.601  1.121 -0.251
 -0.050  1.341  2.282  0.745  1.633  0.944  2.370  0.293  0.895  0.938
  0.199  0.812  1.253  0.590  1.522 -0.685  1.259  0.571  1.579  0.568
  0.381  0.829  0.277  0.656  2.497  1.779  1.922 -0.174  2.132  2.793
  0.102  1.569  1.267  0.490  0.077  1.366  0.056  0.605  0.628  1.650
  0.104  2.194  2.869 -0.171 -0.598  2.134  0.917  0.630  0.209  1.328
  0.368  0.756  2.645  1.161  0.347  0.920  1.256 -0.052  1.474  0.510
  1.386  3.550  1.392 -0.358  1.938  1.727 -0.372  0.911  0.499  0.066
  1.467  1.898  1.145  0.501  2.230  0.212  0.536  1.690  1.086  0.494
 Confidence Level
 0.95
```

## 9.3 Program Results

```
G07EBF Example Program Results

 Location estimator     Confidence Interval

      0.9505            ( 0.5650 , 1.3050 )

 Corresponding Mann-Whitney U statistics

Lower :  2007.00
Upper :  2993.00
```