NAG Fortran Library Routine Document

G01DDF

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

G01DDF calculates Shapiro and Wilk's W statistic and its significance level for testing Normality.

2 Specification

```
SUBROUTINE G01DDF(X, N, CALWTS, A, W, PW, IFAIL)
INTEGER           N, IFAIL
real              X(N), A(N), W, PW
LOGICAL           CALWTS
```

3 Description

This routine calculates Shapiro and Wilk's W statistic and its significance level for any sample size between 3 and 2000. It is an adaptation of the Applied Statistics Algorithm AS 181, see Royston (1982a). The full description of the theory behind this algorithm is given in Royston (1982b).

Given a set of observations $x_1, x_2, \ldots, x_n$ sorted into either ascending or descending order (M01CAF may be used to sort the data) G01DDF calculates the value of Shapiro and Wilk's W statistic defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\bar{x} = \dfrac{1}{n}\sum_{1}^{n} x_i$ is the sample mean and a_i, for $i = 1, 2, \ldots, n$, are a set of 'weights' whose values depend only on the sample size n.

On exit, the values of a_i, for $i = 1, 2, \ldots, n$ are only of interest should the user wish to call the routine again to calculate W and its significance level for a different sample of the same size.

It is recommended that the routine is used in conjunction with G01AHF to give a Normal plot of the data.

4 References

Royston J P (1982a) Algorithm AS181: The W test for Normality *Appl. Statist.* **31** 176–180

Royston J P (1982b) An extension of Shapiro and Wilk's W test for Normality to large samples *Appl. Statist.* **31** 115–124

Royston J P (1986) A remark on AS181: The W test for Normality *Appl. Statist.* **35** 232–234

5 Parameters

1: X(N) – ***real*** array *Input*

On entry: the ordered sample values, x_i; for $i = 1, 2, \ldots, n$.

2: N – INTEGER *Input*

On entry: the sample size, n.

Constraint: $3 \leq \mathrm{N} \leq 2000$.

3: CALWTS – LOGICAL *Input*

On entry: CALWTS must be set to .TRUE. if the user wishes G01DDF to calculate the elements of A.

CALWTS should be set to .FALSE. if the user has saved the values in A from a previous call to G01DDF.

If in doubt, set CALWTS equal to .TRUE..

4: A(N) – ***real*** array *Input/Output*

On entry: if CALWTS has been set to .FALSE. then before entry A must contain the n weights as calculated in a previous call to G01DDF, otherwise A need not be set.

On exit: the n weights required to calculate W.

5: W – ***real*** *Output*

On exit: the value of the statistic, W.

6: PW – ***real*** *Output*

On exit: the significance level of W.

7: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.

On exit: IFAIL $= 0$ unless the routine detects an error (see Section 6).

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, for users 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.**

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, N < 3.

IFAIL $= 2$

On entry, N > 2000.

IFAIL $= 3$

On entry, the elements in X are not in ascending or descending order or are all equal.

7 Accuracy

There may be a loss of significant figures for large n.

8 Further Comments

The time taken by the routine depends roughly linearly on the value of n.

For very small samples the power of the test may not be very high.

The contents of the array A should not be modified between calls to G01DDF for a given sample size, unless CALWTS is reset to .TRUE. before each call of G01DDF.

The Shapiro and Wilk W test is very sensitive to ties. If the data has been rounded the test can be improved by using Sheppard's correction to adjust the sum of squares about the mean. This produces an adjusted value of W,

$$WA = W\frac{\sum(x_{(i)}-\bar{x})^2}{\left\{\sum_{i=1}^{n}(x_{(i)}=\bar{x})^2-\frac{n-1}{12}\omega^2\right\}}$$

where ω is the rounding width. WA can be compared with a standard normal distribution, but a further approximation is given by Royston (1986).

9 Example

A program to test the following 2 samples (each of size 20) for Normality.

Sample Number	**Data**
1	0.11, 7.87, 4.61, 10.14, 7.95, 3.14, 0.46, 4.43, 0.21, 4.75, 0.71, 1.52, 3.24, 0.93, 0.42, 4.97, 9.53, 4.55, 0.47, 6.66
2	1.36, 1.14, 2.92, 2.55, 1.46, 1.06, 5.27, -1.11, 3.48, 1.10, 0.88, -0.51, 1.46, 0.52, 6.20, 1.69, 0.08, 3.67, 2.81, 3.49

The elements of A are calculated only in the first call of G01DDF, and are re-used in the second call.

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.

```
*     G01DDF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NMAX
      PARAMETER        (NMAX=20)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      real             PW, W
      INTEGER          I, IFAIL, J, N
      LOGICAL          CALWTS
*     .. Local Arrays ..
      real             A(NMAX), X(NMAX)
*     .. External Subroutines ..
      EXTERNAL         G01DDF, M01CAF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G01DDF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      CALWTS = .TRUE.
      READ (NIN,*) N
      IF (N.GT.0 .AND. N.LE.NMAX) THEN
         DO 20 J = 1, 2
            READ (NIN,*) (X(I),I=1,N)
            IFAIL = 0
*
            CALL M01CAF(X,1,N,'A',IFAIL)
*
            IFAIL = 0
*
            CALL G01DDF(X,N,CALWTS,A,W,PW,IFAIL)
*
            WRITE (NOUT,*)
            WRITE (NOUT,99999) 'For sample number ', J,
     +        ', value of W statistic = ', W
```

```
            WRITE (NOUT,99998)
     +        '                    Significance level is ', PW
            CALWTS = .FALSE.
   20    CONTINUE
      END IF
      STOP
*
99999 FORMAT (1X,A,I1,A,F7.4)
99998 FORMAT (1X,A,F8.4)
      END
```

9.2 Program Data

```
G01DDF Example Program Data
  20
  0.11  7.87  4.61 10.14  7.95  3.14  0.46  4.43  0.21  4.75
  0.71  1.52  3.24  0.93  0.42  4.97  9.53  4.55  0.47  6.66
  1.36  1.14  2.92  2.55  1.46  1.06  5.27 -1.11  3.48  1.10
  0.88 -0.51  1.46  0.52  6.20  1.69  0.08  3.67  2.81  3.49
```

9.3 Program Results

```
 G01DDF Example Program Results

 For sample number 1, value of W statistic =  0.8992
                      Significance level is   0.0408

 For sample number 2, value of W statistic =  0.9583
                      Significance level is   0.5171
```