

NAG 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
double precision X(N), A(N), W, PW
LOGICAL          CALWTS
```

3 Description

G01DDF calculates Shapiro and Wilk's W statistic and its significance level for any sample size between 3 and 5000. It is an adaptation of the Applied Statistics Algorithm AS R94, see Royston (1995). The full description of the theory behind this algorithm is given in Royston (1992).

Given a set of observations $x_1, x_2, \dots, x_n$ sorted into either ascending or descending order (M01CAF may be used to sort the data) this routine 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 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean and a_i , for $i = 1, 2, \dots, n$, are a set of 'weights' whose values depend only on the sample size n .

On exit, the values of a_i , for $i = 1, 2, \dots, n$, are only of interest should you 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 a Normal ($Q - Q$) plot of the data. Routines G01DAF and G01DBF can be used to obtain the required Normal scores.

4 References

Royston J P (1986) A remark on AS 181: The W test for normality *Appl. Statist.* **35** 232–234

Royston J P (1992) Approximating the Shapiro–Wilk's W test for non-normality *Statistics & Computing* **2** 117–119

Royston J P (1995) A remark on AS R94: A remark on Algorithm AS 181: The W test for normality *Appl. Statist.* **44**(4) 547–551

5 Parameters

- 1: X(N) – **double precision** array *Input*
On entry: the ordered sample values, x_i , for $i = 1, 2, \dots, n$.

- 2: N – INTEGER *Input*
On entry: n , the sample size.
Constraint: $3 \leq N \leq 5000$.
- 3: CALWTS – LOGICAL *Input*
On entry: must be set to .TRUE. if you wish G01DDF to calculate the elements of A.
 CALWTS should be set to .FALSE. if you have saved the values in A from a previous call to G01DDF.
 If in doubt, set CALWTS equal to .TRUE..
- 4: A(N) – *double precision* 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 – *double precision* *Output*
On exit: the value of the statistic, W.
- 6: PW – *double precision* *Output*
On exit: the significance level of W.
- 7: 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.
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, 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.**

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, $N > 5000$.

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 G01DDF 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's 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).

If $N > 5000$, a value for W and PW is returned, but its accuracy may not be acceptable. See Section 4 for more details.

9 Example

A program to test the following two 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

```
*      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 ..
      DOUBLE PRECISION PW, W
      INTEGER          I, IFAIL, J, N
      LOGICAL          CALWTS
*      .. Local Arrays ..
      DOUBLE PRECISION 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 = 1
*
          CALL M01CAF(X,1,N,'A',IFAIL)
*
```

```

      IF (IFAIL.NE.0) THEN
        WRITE (NOUT,*)
        WRITE (NOUT,99999) ' ** M01CAF returned with IFAIL = ',
+         IFAIL
        GO TO 40
      END IF
*
      IFAIL = 1
*
      CALL G01DDF(X,N,CALWTS,A,W,PW,IFAIL)
*
      IF (IFAIL.NE.0) THEN
        WRITE (NOUT,*)
        WRITE (NOUT,99999) ' ** G01DDF returned with IFAIL = ',
+         IFAIL
        GO TO 40
      END IF
*
      WRITE (NOUT,*)
      WRITE (NOUT,99998) 'For sample number ', J,
+      ', value of W statistic = ', W
      WRITE (NOUT,99997)
+      ' Significance level is ', PW
      CALWTS = .FALSE.
20    CONTINUE
      END IF
40  CONTINUE
*
99999 FORMAT (1X,A,I5)
99998 FORMAT (1X,A,I1,A,F7.4)
99997 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 = 0.9005
                      Significance level is 0.0421

```

```

For sample number 2, value of W statistic = 0.9590
                      Significance level is 0.5246

```
