G08CGF – 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

G08CGF computes the test statistic for the χ^2 goodness of fit test for data with a chosen number of class intervals.

2 Specification

```
 SUBROUTINE G08CGF(NCLASS, IFREQ, CINT, DIST, PAR, NPEST, PROB,
1                  CHISQ, P, NDF, EVAL, CHISQI, IFAIL)
 INTEGER           NCLASS, IFREQ(NCLASS), NPEST, NDF, IFAIL
 real              CINT(NCLASS-1), PAR(2), PROB(NCLASS), CHISQ, P,
1                  EVAL(NCLASS), CHISQI(NCLASS)
 CHARACTER*1       DIST
```

3 Description

The χ^2 goodness of fit test performed by G08CGF is used to test the null hypothesis that a random sample arises from a specified distribution against the alternative hypothesis that the sample does not arise from the specified distribution.

Given a sample of size n, denoted by $x_1, x_2, \ldots, x_n$, drawn from a random variable X, and that the data have been grouped into k classes,

$$x \leq c_1,$$
$$c_{i-1} < x \leq c_i, \quad i = 2, 3, \ldots, k-1,$$
$$x > c_{k-1},$$

then the χ^2 goodness of fit test statistic is defined by:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency of the ith class, and E_i is the expected frequency of the ith class.

The expected frequencies are computed as

$$E_i = p_i \times n,$$

where p_i is the probability that X lies in the ith class, that is

$$p_1 = P(X \leq c_1),$$
$$p_i = P(c_{i-1} < X \leq c_i), \quad i = 2, 3, \ldots, k-1,$$
$$p_k = P(X > c_{k-1}).$$

These probabilities are either taken from a common probability distribution or are supplied by the user. The available probability distributions within this routine are:

Normal distribution with mean μ, variance σ^2;

uniform distribution on the interval $[a, b]$;

exponential distribution with probability density function (pdf) $= \lambda e^{-\lambda x}$;

χ^2 distribution with f degrees of freedom; and

gamma distribution with pdf $= \dfrac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$.

The user must supply the frequencies and classes. Given a set of data and classes the frequencies may be calculated using G01AEF.

G08CGF returns the χ^2 test statistic, X^2, together with its degrees of freedom and the upper tail probability from the χ^2 distribution associated with the test statistic. Note that the use of the χ^2 distribution as an approximation to the distribution of the test statistic improves as the expected values in each class increase.

4 References

[1] Conover W J (1980) *Practical Nonparametric Statistics* Wiley

[2] Kendall M G and Stuart A (1973) *The Advanced Theory of Statistics (Volume 2)* Griffin (3rd Edition)

[3] Siegel S (1956) *Non-parametric Statistics for the Behavioral Sciences* McGraw–Hill

5 Parameters

1: NCLASS — INTEGER *Input*

On entry: the number of classes, k, into which the data is divided.

Constraint: NCLASS ≥ 2.

2: IFREQ(NCLASS) — INTEGER array *Input*

On entry: IFREQ(i) must specify the frequency of the ith class, O_i, for $i = 1, 2, \ldots, k$.

Constraint: IFREQ(i) ≥ 0, for $i = 1, 2, \ldots, k$.

3: CINT(NCLASS-1) — ***real*** array *Input*

On entry: CINT(i) must specify the upper boundary value for the ith class, for $i = 1, 2, \ldots, k - 1$.

Constraint: CINT(1) < CINT(2) < ... < CINT(NCLASS – 1). For the exponential, gamma and χ^2 distributions CINT(1) $\geq$ 0.0.

4: DIST — CHARACTER*1 *Input*

On entry: indicates for which distribution the test is to be carried out;

If DIST = ’N’, the Normal distribution is used.

If DIST = ’U’, the uniform distribution is used.

If DIST = ’E’, the exponential distribution is used.

If DIST = ’C’, the χ^2 distribution is used.

If DIST = ’G’, the gamma distribution is used.

If DIST = ’A’, the user must supply the class probabilities in the array PROB.

Constraint: DIST = ’N’, ’U’, ’E’, ’C’, ’G’ or ’A’.

5: PAR(2) — ***real*** array *Input*

On entry: PAR must contain the parameters of the distribution which is being tested. If the user supplies the probabilities (that is, DIST = 'A') the array PAR is not referenced.

If a Normal distribution is used then PAR(1) and PAR(2) must contain the mean, μ, and the variance, σ^2, respectively.

If a uniform distribution is used then PAR(1) and PAR(2) must contain the boundaries a and b respectively.

If an exponential distribution is used then PAR(1) must contain the parameter λ. PAR(2) is not used.

If a χ^2 distribution is used then PAR(1) must contain the number of degrees of freedom. PAR(2) is not used.

If a gamma distribution is used PAR(1) and PAR(2) must contain the parameters α and β respectively.

Constraints:

if DIST = 'N', PAR(2) > 0.0,
if DIST = 'U', PAR(1) < PAR(2), PAR(1) ≤ CINT(1),
PAR(2) ≥ CINT(NCLASS−1),
if DIST = 'E', PAR(1) > 0.0,
if DIST = 'C', PAR(1) > 0.0,
if DIST = 'G', PAR(1), PAR(2) > 0.0.

6: NPEST — INTEGER *Input*

On entry: the number of estimated parameters of the distribution.

Constraint: 0 ≤ NPEST < NCLASS − 1.

7: PROB(NCLASS) — ***real*** array *Input*

On entry: if the user is supplying the probability distribution (that is, DIST = 'A') then PROB(i) must contain the probability that X lies in the ith class.

If DIST ≠ 'A', PROB is not referenced.

Constraints: if DIST = 'A' then PROB(i) > 0.0, for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k}$PROB(i) = 1.0.

8: CHISQ — ***real*** *Output*

On exit: the test statistic, X^2, for the χ^2 goodness of fit test.

9: P — ***real*** *Output*

On exit: the upper tail probability from the χ^2 distribution associated with the test statistic, X^2, and the number of degrees of freedom.

10: NDF — INTEGER *Output*

On exit: contains (NCLASS – 1 – NPEST), the degrees of freedom associated with the test.

11: EVAL(NCLASS) — ***real*** array *Output*

On exit: EVAL(i) contains the expected frequency for the ith class, E_i, for $i = 1, 2, \ldots, k$.

12: CHISQI(NCLASS) — ***real*** array *Output*

On exit: CHISQI(i) contains the contribution from the ith class to the test statistic, that is $(O_i - E_i)^2/E_i$, for $i = 1, 2, \ldots, k$.

13: 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 or gives a warning (see Section 6).

For this routine, because the values of output parameters may be useful even if IFAIL $\neq 0$ on exit, users are recommended to set IFAIL to −1 before entry. **It is then 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 specified by the routine:

IFAIL = 1

On entry, NCLASS < 2.

IFAIL = 2

On entry, DIST is invalid.

IFAIL = 3

On entry, NPEST < 0,

or NPEST $\geq$ NCLASS − 1.

IFAIL = 4

On entry, IFREQ(i) < 0.0 for some i, for $i = 1, 2,.$

IFAIL = 5

On entry, the elements of CINT are not in ascending order. That is CINT(i) $\leq$ CINT($i-1$) for some i, for $i = 2, 3, \ldots, k-1$.

IFAIL = 6

On entry, DIST = 'E', 'C' or 'G' and CINT(1) < 0.0. No negative class boundary values are valid for the exponential, gamma or χ^2 distributions.

IFAIL = 7

On entry, the values provided in PAR are invalid.

IFAIL = 8

On entry, with DIST = 'A', PROB(i) $\leq$ 0.0 for some i, for $i = 1, 2, \ldots, k$,

or $\sum_{i=1}^{k}$PROB(i) $\neq$ 1.0.

IFAIL = 9

An expected frequency is equal to zero when the observed frequency was not.

IFAIL = 10

This is a warning that expected values for certain classes are less than 1.0. This implies that we cannot be confident that the χ^2 distribution is a good approximation to the distribution of the test statistic.

IFAIL = 11

The solution obtained when calculating the probability for a certain class for the gamma or χ^2 distribution did not converge in 600 iterations. The solution may be an adequate approximation.

7 Accuracy

The computations are believed to be stable.

8 Further Comments

The time taken by the routine is dependent both on the distribution chosen and on the number of classes, k.

9 Example

The example program applies the χ^2 goodness of fit test to test whether there is evidence to suggest that a sample of 100 observations generated by G05DAF do not arise from a uniform distribution $U(0,1)$. The class intervals are calculated such that the interval (0,1) is divided into 5 equal classes. The frequencies for each class are calculated using G01AEF.

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.

```
*     G08CGF Example Program Text
*     Mark 15 Revised.  NAG Copyright 1991.
*     .. Parameters ..
      INTEGER          NIN, NOUT, NMAX, NCLMAX
      PARAMETER        (NIN=5,NOUT=6,NMAX=100,NCLMAX=10)
*     .. Local Scalars ..
      real             CHISQ, P, XMAX, XMIN
      INTEGER          I, ICLASS, IFAIL, NPEST, NCLASS, N, NDF
      CHARACTER*1      CDIST
*     .. Local Arrays ..
      real             CHISQI(NCLMAX), CINT(NCLMAX), EVAL(NCLMAX),
     +                 PAR(2), PROB(NCLMAX), X(NMAX)
      INTEGER          IFREQ(NCLMAX)
*     .. External Subroutines ..
      EXTERNAL         G01AEF, G05CBF, G05FAF, G08CGF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G08CGF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, NCLASS, CDIST
      READ (NIN,*) (PAR(I),I=1,2)
      NPEST = 0
*
*     Generate random numbers from a uniform distribution
      CALL G05CBF(0)
*
      CALL G05FAF(PAR(1),PAR(2),N,X)
      ICLASS = 0
*
*     Determine suitable intervals
      IF (CDIST.EQ.'U' .OR. CDIST.EQ.'u') THEN
         ICLASS = 1
         CINT(1) = PAR(1) + (PAR(2)-PAR(1))/NCLASS
         DO 20 I = 2, NCLASS - 1
            CINT(I) = CINT(I-1) + (PAR(2)-PAR(1))/NCLASS
   20    CONTINUE
```

```
      END IF
      IFAIL = 0
*
      CALL G01AEF(N,NCLASS,X,ICLASS,CINT,IFREQ,XMIN,XMAX,IFAIL)
*
      IFAIL = 0
*
      CALL G08CGF(NCLASS,IFREQ,CINT,CDIST,PAR,NPEST,PROB,CHISQ,P,NDF,
     +            EVAL,CHISQI,IFAIL)
*
      IF (IFAIL.NE.0) WRITE (NOUT,99999) '** IFAIL = ', IFAIL
      WRITE (NOUT,*)
      WRITE (NOUT,99998) 'Chi-squared test statistic   = ', CHISQ
      WRITE (NOUT,99997) 'Degrees of freedom.          = ', NDF
      WRITE (NOUT,99998) 'Significance level           = ', P
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'The contributions to the test statistic are :-'
      DO 40 I = 1, NCLASS
         WRITE (NOUT,99996) CHISQI(I)
   40 CONTINUE
      STOP
*
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,A,F10.4)
99997 FORMAT (1X,A,I5)
99996 FORMAT (1X,F10.4)
      END
```

9.2 Program Data

```
G08CGF Example Program Data.
100 5 'U'      :N  K2  CDIST
0.0 1.0        :PAR(1) PAR(2)
```

9.3 Program Results

```
 G08CGF Example Program Results

 Chi-squared test statistic   =     3.3000
 Degrees of freedom.          =     4
 Significance level           =     0.5089

 The contributions to the test statistic are :-
     1.8000
     0.8000
     0.2000
     0.0500
     0.4500
```