NAG Library Routine Document

G08CGF

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, CB, DIST, PAR, NPEST, PROB, CHISQ, P,         &
                   NDF, EVAL, CHISQI, IFAIL)

INTEGER            NCLASS, IFREQ(NCLASS), NPEST, NDF, IFAIL
REAL (KIND=nag_wp) CB(NCLASS-1), PAR(2), PROB(NCLASS), CHISQ, P,                &
                   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 has been grouped into k classes,

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

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

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

These probabilities are either taken from a common probability distribution or are supplied by you. 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 $(\text{pdf}) = \lambda e^{-\lambda x}$;

χ^2-distribution with f degrees of freedom; and

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

You 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

Conover W J (1980) *Practical Nonparametric Statistics* Wiley

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

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

5 Parameters

1: NCLASS – INTEGER *Input*

On entry: k, the number of classes 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$(i) \geq 0$, for $i = 1, 2, \ldots, k$.

3: CB(NCLASS – 1) – REAL (KIND=nag_wp) array *Input*

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

Constraint: $\mathrm{CB}(1) < \mathrm{CB}(2) < \cdots < \mathrm{CB}(\mathrm{NCLASS} - 1)$. For the exponential, gamma and χ^2-distributions $\mathrm{CB}(1) \geq 0.0$.

4: DIST – CHARACTER(1) *Input*

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

DIST = 'N'
The Normal distribution is used.

DIST = 'U'
The uniform distribution is used.

DIST = 'E'
The exponential distribution is used.

DIST = 'C'
The χ^2-distribution is used.

DIST = 'G'
The gamma distribution is used.

DIST = 'A'
You must supply the class probabilities in the array PROB.

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

5: PAR(2) – REAL (KIND=nag_wp) array *Input*

On entry: must contain the parameters of the distribution which is being tested. If you supply the probabilities (i.e., 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) and PAR(1) $\leq$ CB(1) and PAR(2) $\geq$ CB(NCLASS $-$ 1);
if DIST = 'E', PAR(1) > 0.0;
if DIST = 'C', PAR(1) > 0.0;
if DIST = 'G', PAR(1) > 0.0 and PAR(2) > 0.0.

6: NPEST – INTEGER *Input*

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

Constraint: $0 \leq$ NPEST $<$ NCLASS $-$ 1.

7: PROB(NCLASS) – REAL (KIND=nag_wp) array *Input*

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

If DIST $\neq$ 'A', PROB is not referenced.

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

8: CHISQ – REAL (KIND=nag_wp) *Output*

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

9: P – REAL (KIND=nag_wp) *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 (KIND=nag_wp) 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 (KIND=nag_wp) 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. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

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, because for this routine the values of the output parameters may be useful even if IFAIL $\neq 0$ on exit, the recommended value is -1. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (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).

Note: G08CGF may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected 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 $i = 1, 2, \ldots, k$.

IFAIL $= 5$

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

IFAIL $= 6$

On entry, DIST = 'E', 'C' or 'G' and CB$(1) < 0.0$. 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 G08CGF is dependent both on the distribution chosen and on the number of classes, k.

9 Example

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

9.1 Program Text

```
    Program g08cgfe

!     G08CGF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g01aef, g08cgf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: chisq, p, xmax, xmin
      Integer                          :: iclass, ifail, n, nclass, ndf, npar, &
                                          npest
      Character (1)                    :: dist
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: cb(:), chisqi(:), eval(:), prob(:),  &
                                          x(:)
      Real (Kind=nag_wp)               :: par(2)
      Integer, Allocatable             :: ifreq(:)
!     .. Executable Statements ..
      Write (nout,*) 'G08CGF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in problem size
      Read (nin,*) n

!     Read in class information
      Read (nin,*) nclass, iclass

      Allocate (x(n),cb(nclass),ifreq(nclass),prob(nclass),eval(nclass), &
        chisqi(nclass))
```

```
!     Read in data
      Read (nin,*) x(1:n)

!     Read in the class boundaries, if supplied
      If (iclass==1) Then
        Read (nin,*) cb(1:(nclass-1))
      End If

!     Read in information on the distribution to test against
      Read (nin,*) dist, npest

      Select Case (dist)
      Case (’A’,’a’)
        npar = 0
      Case (’E’,’e’,’C’,’c’)
        npar = 1
      Case Default
        npar = 2
      End Select

!     Read in the distribution parameters or probabilities
      If (npar==0) Then
        Read (nin,*) prob(1:nclass)
      Else
        Read (nin,*) par(1:npar)
      End If

!     Produce frequency table for data
      ifail = 0
      Call g01aef(n,nclass,x,iclass,cb,ifreq,xmin,xmax,ifail)

!     Perform chi-squared test
      ifail = -1
      Call g08cgf(nclass,ifreq,cb,dist,par,npest,prob,chisq,p,ndf,eval,chisqi, &
        ifail)
      If (ifail/=0) Then
        If (ifail<=9) Then
          Go To 100
        End If
      End If

!     Display results
      Write (nout,99999) ’Chi-squared test statistic   = ’, chisq
      Write (nout,99998) ’Degrees of freedom.          = ’, ndf
      Write (nout,99999) ’Significance level           = ’, p
      Write (nout,*)
      Write (nout,*) ’The contributions to the test statistic are :-’
      Write (nout,99997) chisqi(1:nclass)

100   Continue

99999 Format (1X,A,F10.4)
99998 Format (1X,A,I5)
99997 Format (1X,F10.4)
    End Program g08cgfe
```

9.2 Program Data

```
G08CGF Example Program Data
100                                               :: N
5   1                                             :: NCLASS,ICLASS
0.59 0.23 0.76 0.96 0.20 0.91 0.29 0.22 0.36 0.81
0.91 0.80 0.17 0.82 0.07 0.74 0.15 0.91 0.26 0.98
0.59 0.34 0.28 0.95 0.33 0.42 0.72 0.35 0.86 0.22
0.15 0.39 0.32 0.82 0.13 0.48 0.46 0.74 0.99 0.26
0.04 0.21 0.04 0.24 0.56 0.36 0.48 0.53 1.00 0.58
0.50 0.41 0.03 0.38 0.89 0.40 0.66 0.79 0.34 0.94
0.49 0.12 0.24 0.05 1.00 0.29 0.67 0.29 0.75 0.81
0.45 0.21 0.51 0.68 0.78 0.20 0.23 0.57 0.25 0.48
```

```
0.96 0.33 0.48 0.55 0.04 0.48 0.42 0.11 0.38 0.73
0.91 0.45 0.59 0.97 0.27 0.27 0.25 0.99 0.99 0.80 :: End of X
0.2 0.4 0.6 0.8                                   :: CB
'U'  0                                            :: DIST,NPEST
0.0 1.0                                           :: PAR
```

9.3 Program Results

```
 G08CGF Example Program Results

 Chi-squared test statistic   =    14.2000
 Degrees of freedom.          =     4
 Significance level           =     0.0067

 The contributions to the test statistic are :-
     3.2000
     6.0500
     0.4500
     4.0500
     0.4500
```