# NAG Toolbox

# nag_nonpar_test_chisq (g08cg)

## 1 Purpose

nag_nonpar_test_chisq (g08cg) computes the test statistic for the $\chi^2$ goodness-of-fit test for data with a chosen number of class intervals.

## 2 Syntax

```
[chisq, p, ndf, eval, chisqi, ifail] = nag_nonpar_test_chisq(ifreq, cb, dist, par, npest, prob, 'nclass', nclass)
[chisq, p, ndf, eval, chisqi, ifail] = g08cg(ifreq, cb, dist, par, npest, prob, 'nclass', nclass)
```

## 3 Description

The $\chi^2$ goodness-of-fit test performed by nag_nonpar_test_chisq (g08cg) 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,

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

then the $\chi^2$ 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 $i$th class, and $E_i$ is the expected frequency of the $i$th class.

The expected frequencies are computed as

$$E_i = p_i \times n,$$

where $p_i$ is the probability that $X$ lies in the $i$th class, that is

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

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

Normal distribution with mean $\mu$, variance $\sigma^2$;

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

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

$\chi^2$-distribution with $f$ degrees of freedom; and

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

You must supply the frequencies and classes. Given a set of data and classes the frequencies may be calculated using nag_stat_frequency_table (g01ae).

nag_nonpar_test_chisq (g08cg) returns the $\chi^2$ test statistic, $X^2$, together with its degrees of freedom and the upper tail probability from the $\chi^2$-distribution associated with the test statistic. Note that the use of the $\chi^2$-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

## 5.1 Compulsory Input Parameters

1: **ifreq**(**nclass**) – INTEGER array

**ifreq**($i$) must specify the frequency of the $i$th class, $O_i$, for $i = 1, 2, \ldots, k$.

*Constraint*: **ifreq**($i$) $\geq 0$, for $i = 1, 2, \ldots, k$.

2: **cb**(**nclass** $-$ **1**) – REAL (KIND=nag_wp) array

**cb**($i$) must specify the upper boundary value for the $i$th class, for $i = 1, 2, \ldots, k-1$.

*Constraint*: **cb**(1) $<$ **cb**(2) $< \cdots <$ **cb**(**nclass** $-$ 1). For the exponential, gamma and $\chi^2$-distributions **cb**(1) $\geq 0.0$.

3: **dist** – CHARACTER(1)

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 $\chi^2$-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'.

4: **par**(**2**) – REAL (KIND=nag_wp) array

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, $\mu$, and the variance, $\sigma^2$, 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 $\lambda$. **par**(2) is not used.

If a $\chi^2$-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 $\alpha$ and $\beta$ 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**$(2) > 0.0$.

5: **npest** – INTEGER

The number of estimated parameters of the distribution.

*Constraint*: $0 \leq$ **npest** $<$ **nclass** $- 1$.

6: **prob**$($**nclass**$)$ – REAL (KIND=nag_wp) array

If you are supplying the probability distribution (i.e., **dist** = 'A') then **prob**$(i)$ must contain the probability that $X$ lies in the $i$th 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 $i = 1, 2, \ldots, k$.

## 5.2 Optional Input Parameters

1: **nclass** – INTEGER

*Default*: the dimension of the arrays **ifreq**, **prob**. (An error is raised if these dimensions are not equal.)

$k$, the number of classes into which the data is divided.

*Constraint*: **nclass** $\geq 2$.

## 5.3 Output Parameters

1: **chisq** – REAL (KIND=nag_wp)

The test statistic, $X^2$, for the $\chi^2$ goodness-of-fit test.

2: **p** – REAL (KIND=nag_wp)

The upper tail probability from the $\chi^2$-distribution associated with the test statistic, $X^2$, and the number of degrees of freedom.

3: **ndf** – INTEGER

Contains (**nclass** $- 1 -$ **npest**), the degrees of freedom associated with the test.

4: **eval**$($**nclass**$)$ – REAL (KIND=nag_wp) array

**eval**$(i)$ contains the expected frequency for the $i$th class, $E_i$, for $i = 1, 2, \ldots, k$.

5: **chisqi**$($**nclass**$)$ – REAL (KIND=nag_wp) array

**chisqi**$(i)$ contains the contribution from the $i$th class to the test statistic, that is, $(O_i - E_i)^2/E_i$, for $i = 1, 2, \ldots, k$.

6: **ifail** – INTEGER

**ifail** $= 0$ unless the function detects an error (see Section 5).

## 6 Error Indicators and Warnings

**Note**: nag_nonpar_test_chisq (g08cg) may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the function:

**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 $\chi^2$-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** $= 10$ (*warning*)

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

**ifail** $= 11$ (*warning*)

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

**ifail** $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

**ifail** $= -399$

Your licence key may have expired or may not have been installed correctly.

**ifail** $= -999$

Dynamic memory allocation failed.

## 7 Accuracy

The computations are believed to be stable.

## 8 Further Comments

The time taken by nag_nonpar_test_chisq (g08cg) is dependent both on the distribution chosen and on the number of classes, $k$.

## 9 Example

This example applies the $\chi^2$ 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 nag_stat_frequency_table (g01ae).

### 9.1 Program Text

```
function g08cg_example

fprintf('g08cg example results\n\n');

x = [ 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];

cb     = [0.2;     0.4;     0.6;     0.8;    1.0 ];
nclass = nag_int(5);

% Produce frequency table
[~, ifreq, ~, ~, ifail] = ...
  g01ae( ...
         nclass, x, 'cb', cb);

% Test parameters
dist   = 'Uniform';
npest  = nag_int(0);
par    = [0;  1];
prob   = zeros(nclass,1);

% Perform Chi^2 test
[chisq, p, ndf, eval, chisqi, ifail] = ...
  g08cg( ...
         ifreq, cb, dist, par, npest, prob, 'nclass', nclass);
```

```
fprintf('Chi-squared test statistic   = %10.4f\n', chisq);
fprintf('Degrees of freedom.          = %5d\n', ndf);
fprintf('Significance level           = %10.4f\n\n', p);
fprintf('The contributions to the test statistic are :-\n');
disp(chisqi');
```

## 9.2 Program Results

```
     g08cg example 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
```