NAG Toolbox for MATLAB

g08cg

1 Purpose

g08cg computes the test statistic for the χ^2 goodness-of-fit test for data with a chosen number of class intervals.

2 Syntax

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

3 Description

The χ^2 goodness-of-fit test performed by 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,

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

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 function 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 g01ae.

g08cg 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

5.1 Compulsory Input Parameters

1: **ifreq**(**nclass**) – **int32 array**

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

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

2: **cb**(**nclass** $-$ **1**) – **double array**

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

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

3: **dist** – **string**

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'.

4: **par**(**2**) – **double 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, 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**$(2) > 0.0$.

5: **npest – int32 scalar**

The number of estimated parameters of the distribution.

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

6: **prob**(**nclass**) **– double array**

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 $i = 1, 2, \ldots, k$.

5.2 Optional Input Parameters

1: **nclass – int32 scalar**

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

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

Constraint: **nclass** ≥ 2.

5.3 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **chisq – double scalar**

The test statistic, X^2, for the χ^2 goodness-of-fit test.

2: **p – double scalar**

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

3: **ndf – int32 scalar**

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

4: **eval**(**nclass**) **– double array**

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

5: **chisqi(nclass) – double array**

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

6: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

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

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 g08cg is dependent both on the distribution chosen and on the number of classes, k.

9 Example

```
ifreq = [int32(26);
     int32(16);
     int32(22);
     int32(19);
     int32(17)];
cb = [0.2;
     0.4;
     0.6;
     0.7999999999999999];
dist = 'U';
par = [0;
     1];
npest = int32(0);
prob = [0;
     0;
     0;
     4.878438904751203e+199;
     5.495816452771857e+222];
[chisq, p, ndf, eval, chisqi, ifail] = g08cg(ifreq, cb, dist, par, npest,
prob)
```

```
chisq =
    3.3000
p =
    0.5089
ndf =
           4
eval =
   20.0000
   20.0000
   20.0000
   20.0000
   20.0000
chisqi =
    1.8000
    0.8000
    0.2000
    0.0500
    0.4500
ifail =
           0
```