NAG Toolbox for MATLAB

g08cb

1 Purpose

g08cb performs the one-sample Kolmogorov–Smirnov test, using one of the standard distributions provided.

2 Syntax

```
[par, d, z, p, sx, ifail] = g08cb(x, dist, par, estima, ntype, 'n', n)
```

3 Description

The data consists of a single sample of n observations denoted by $x_1, x_2, \ldots, x_n$. Let $S_n\left(x_{(i)}\right)$ and $F_0\left(x_{(i)}\right)$ represent the sample cumulative distribution function and the theoretical (null) cumulative distribution function respectively at the point $x_{(i)}$, where $x_{(i)}$ is the ith smallest sample observation.

The Kolmogorov–Smirnov test provides a test of the null hypothesis H_0: the data are a random sample of observations from a theoretical distribution specified by you against one of the following alternative hypotheses

(i) H_1 : the data cannot be considered to be a random sample from the specified null distribution.

(ii) H_2 : the data arise from a distribution which dominates the specified null distribution. In practical terms, this would be demonstrated if the values of the sample cumulative distribution function $S_n(x)$ tended to exceed the corresponding values of the theoretical cumulative distribution function $F_0(x)$.

(iii) H_3 : the data arise from a distribution which is dominated by the specified null distribution. In practical terms, this would be demonstrated if the values of the theoretical cumulative distribution function $F_0(x)$ tended to exceed the corresponding values of the sample cumulative distribution function $S_n(x)$.

One of the following test statistics is computed depending on the particular alternative null hypothesis specified (see the description of the parameter **ntype** in Section 5).

For the alternative hypothesis H_1 :

> D_n – the largest absolute deviation between the sample cumulative distribution function and the theoretical cumulative distribution function. Formally $D_n = \max\left\{D_n^+, D_n^-\right\}$.

For the alternative hypothesis H_2 :

> D_n^+ – the largest positive deviation between the sample cumulative distribution function and the theoretical cumulative distribution function. Formally $D_n^+ = \max\left\{S_n\left(x_{(i)}\right) - F_0\left(x_{(i)}\right), 0\right\}$ for both discrete and continuous null distributions.

For the alternative hypothesis H_3 :

> D_n^- – the largest positive deviation between the theoretical cumulative distribution function and the sample cumulative distribution function. Formally if the null distribution is discrete then $D_n^- = \max\left\{F_0\left(x_{(i)}\right) - S_n\left(x_{(i)}\right), 0\right\}$ and if the null distribution is continuous then $D_n^- = \max\left\{F_0\left(x_{(i)}\right) - S_n\left(x_{(i-1)}\right), 0\right\}$.

The standardized statistic $Z = D \times \sqrt{n}$ is also computed, where D may be D_n, D_n^+ or D_n^- depending on the choice of the alternative hypothesis. This is the standardized value of D with no correction for continuity applied and the distribution of Z converges asymptotically to a limiting distribution, first derived by Kolmogorov 1933, and then tabulated by Smirnov 1948. The asymptotic distributions for the one-sided statistics were obtained by Smirnov 1933.

The probability, under the null hypothesis, of obtaining a value of the test statistic as extreme as that observed, is computed. If $n \leq 100$ an exact method given by Conover 1980 is used. Note that the method used is only exact for continuous theoretical distributions and does not include Conover's modification for discrete distributions. This method computes the one-sided probabilities. The two-sided probabilities are estimated by doubling the one-sided probability. This is a good estimate for small p, that is $p \leq 0.10$, but it becomes very poor for larger p. If $n > 100$ then p is computed using the Kolmogorov–Smirnov limiting distributions; see Feller 1948, Kendall and Stuart 1973, Kolmogorov 1933, Smirnov 1933 and Smirnov 1948.

4 References

Conover W J 1980 *Practical Nonparametric Statistics* Wiley

Feller W 1948 On the Kolmogorov–Smirnov limit theorems for empirical distributions *Ann. Math. Statist.* **19** 179–181

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

Kolmogorov A N 1933 Sulla determinazione empirica di una legge di distribuzione *Giornale dell' Istituto Italiano degli Attuari* **4** 83–91

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

Smirnov N 1933 Estimate of deviation between empirical distribution functions in two independent samples *Bull. Moscow Univ.* **2 (2)** 3–16

Smirnov N 1948 Table for estimating the goodness of fit of empirical distributions *Ann. Math. Statist.* **19** 279–281

5 Parameters

5.1 Compulsory Input Parameters

1: **x(n) – double array**

The sample observations $x_1, x_2, \ldots, x_n$.

Constraint: the sample observations supplied must be consistent, in the usual manner, with the null distribution chosen, as specified by the parameters **dist** and **par**. For further details see Section 8.

2: **dist – string**

The theoretical (null) distribution from which it is suspected the data may arise.

dist = 'U'

The uniform distribution over $(a,b) - U(a,b)$.

dist = 'N'

The Normal distribution with mean μ and variance $\sigma^2 - \mathbf{n}(\mu,\sigma^2)$.

dist = 'G'

The gamma distribution with shape parameter α and scale parameter β, where the mean $= \alpha\beta$.

dist = 'BE'

The beta distribution with shape parameters α and β, where the mean $= \alpha/(\alpha+\beta)$.

dist = 'BI'

The binomial distribution with the number of trials, m, and the probability of a success, p.

dist = 'E'

The exponential distribution with parameter λ, where the mean $= 1/\lambda$.

dist = 'P'

The Poisson distribution with parameter μ, where the mean $= \mu$.

Any number of characters may be supplied as the actual argument, however only the characters, maximum 2, required to uniquely identify the distribution are referenced.

3: **par**(**2**) **– double array**

If **estima** = 'S', **par** must contain the known values of the parameter(s) of the null distribution as follows.

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

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 gamma distribution is used then **par**(1) and **par**(2) must contain the parameters α and β respectively.

If a beta distribution is used then **par**(1) and **par**(2) must contain the parameters α and β respectively.

If a binomial distribution is used then **par**(1) and **par**(2) must contain the parameters m and p respectively.

If a exponential distribution is used then **par**(1) must contain the parameter λ.

If a poisson distribution is used then **par**(1) must contain the parameter μ.

If **estima** = 'E', **par** need not be set except when the null distribution requested is the binomial distribution in which case **par**(1) must contain the parameter m.

Constraints:

if **dist** = 'U', **par**(1) < **par**(2);
if **dist** = 'N', **par**(2) > 0.0;
if **dist** = 'G', **par**(1) > 0.0 and **par**(2) > 0.0;
if **dist** = 'BE', **par**(1) > 0.0 and **par**(2) > 0.0 and **par**(1) $\leq 10^6$ and **par**(2) $\leq 10^6$;
if **dist** = 'BI', **par**(1) $\geq$ 1.0 and 0.0 < **par**(2) < 1.0 and
par(1) × **par**(2) × (1.0 – **par**(2)) $\leq 10^6$ and **par**(1) $< 1/\epsilon$ where
ϵ = ***machine precision***, see x02aj;
if **dist** = 'E', **par**(1) > 0.0;
if **dist** = 'P', **par**(1) > 0.0 and **par**(1) $\leq 10^6$.

4: **estima – string**

Must specify whether values of the parameters of the null distribution are known or are to be estimated from the data.

estima = 'S'

> Values of the parameters will be supplied in the array **par** described above.

estima = 'E'

> Parameters are to be estimated from the data except when the null distribution requested is the binomial distribution in which case the first parameter, m, must be supplied in **par**(1) and only the second parameter, p, is estimated from the data.

Constraint: **estima** = 'S' or 'E'.

5: **ntype – int32 scalar**

The test statistic to be calculated, i.e., the choice of alternative hypothesis.

ntype $= 1$

> Computes D_n, to test H_0 against H_1.

ntype $= 2$

> Computes D_n^+, to test H_0 against H_2.

ntype $= 3$

> Computes D_n^-, to test H_0 against H_3.

Constraint: **ntype** $= 1$, 2 or 3.

5.2 Optional Input Parameters

1: **n – int32 scalar**

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

n, the number of observations in the sample.

Constraint: $\mathbf{n} \geq 3$.

5.3 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **par(2) – double array**

If **estima** = 'S', **par** is unchanged.

If **estima** = 'E' then **par**(1) and **par**(2) are set to values as estimated from the data.

2: **d – double scalar**

The Kolmogorov–Smirnov test statistic (D_n, D_n^+ or D_n^- according to the value of **ntype**).

3: **z – double scalar**

A standardized value, Z, of the test statistic, D, without any continuity correction applied.

4: **p – double scalar**

The probability, p, associated with the observed value of D, where D may be D_n, D_n^+ or D_n^- depending on the value of **ntype** (see Section 3).

5: **sx(n) – double array**

The sample observations, $x_1, x_2, \ldots, x_n$, sorted in ascending order.

6: **ifail – int32 scalar**

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

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, $\mathbf{n} < 3$.

ifail $= 2$

On entry, an invalid code for **dist** has been specified.

ifail $= 3$

On entry, **ntype** $\neq 1$, 2 or 3.

ifail $= 4$

On entry, **estima** $\neq$ 'S' or 'E'.

ifail $= 5$

On entry, the parameters supplied for the specified null distribution are out of range (see Section 5). Apart from a check on the first parameter for the binomial distribution (**dist** = 'BI') this error will only occur if **estima** = 'S'.

ifail $= 6$

The data supplied in **x** could not arise from the chosen null distribution, as specified by the parameters **dist** and **par**. For further details see Section 8.

ifail $= 7$

The whole sample is constant, i.e., the variance is zero. This error may only occur if (**dist** = 'U', 'N', 'G' or 'BE') and **estima** = 'E'.

ifail $= 8$

The variance of the binomial distribution (**dist** = 'BI') is too large. That is, $mp(1-p) > 1000000$.

ifail $= 9$

When **dist** = 'G', in the computation of the incomplete gamma function by s14ba the convergence of the Taylor series or Legendre continued fraction fails within 600 iterations. This is an unlikely error exit.

7 Accuracy

The approximation for p, given when $n > 100$, has a relative error of at most 2.5% for most cases. The two-sided probability is approximated by doubling the one-sided probability. This is only good for small p, i.e., $p < 0.10$, but very poor for large p. The error is always on the conservative side, that is the tail probability, p, is over estimated.

8 Further Comments

The time taken by g08cb increases with n until $n > 100$ at which point it drops and then increases slowly with n. The time may also depend on the choice of null distribution and on whether or not the parameters are to be estimated.

The data supplied in the parameter **x** must be consistent with the chosen null distribution as follows.

When **dist** = 'U', then $\mathbf{par}(1) \leq x_i \leq \mathbf{par}(2)$ for $i = 1, 2, \ldots, n$.

When **dist** = 'N', then there are no constraints on the x_i.

When **dist** = 'G', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$.

When **dist** = 'BE' then $0.0 \leq x_i \leq 1.0$, for $i = 1, 2, \ldots, n$.

When **dist** = 'BI', then $0.0 \leq x_i \leq \mathbf{par}(1)$, for $i = 1, 2, \ldots, n$.

When **dist** = 'E', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$.

When **dist** = 'P', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$.

9 Example

```
x = [0.01;
     0.3;
     0.2;
     0.9;
     1.2;
     0.09;
     1.3;
     0.18;
     0.9;
     0.48;
     1.98;
     0.03;
     0.5;
     0.07000000000000001;
     0.7;
     0.6;
     0.95;
     1;
     0.31;
     1.45;
     1.04;
     1.25;
     0.15;
     0.75;
     0.85;
     0.22;
     1.56;
     0.8100000000000001;
     0.57;
     0.55];
dist = 'Uniform';
par = [0;
     2];
estima = 'Supplied';
ntype = int32(1);
[parOut, d, z, p, sx, ifail] = g08cb(x, dist, par, estima, ntype)
```

```
parOut =
     0
     2
d =
    0.2800
z =
    1.5336
```

```
p =
    0.0143
sx =
    0.0100
    0.0300
    0.0700
    0.0900
    0.1500
    0.1800
    0.2000
    0.2200
    0.3000
    0.3100
    0.4800
    0.5000
    0.5500
    0.5700
    0.6000
    0.7000
    0.7500
    0.8100
    0.8500
    0.9000
    0.9000
    0.9500
    1.0000
    1.0400
    1.2000
    1.2500
    1.3000
    1.4500
    1.5600
    1.9800
ifail =
           0
```