NAG Toolbox for MATLAB

g07ea

1 Purpose

g07ea computes a rank based (nonparametric) estimate and confidence interval for the location parameter of a single population.

2 Syntax

```
[theta, thetal, thetau, estcl, wlower, wupper, ifail] = g07ea(method, x,
clevel, 'n', n)
```

3 Description

Consider a vector of independent observations, $x = (x_1, x_2, \dots, x_n)^{\mathrm{T}}$ with unknown common symmetric density $f(x_i - \theta)$. g07ea computes the Hodges–Lehmann location estimator (see Lehmann 1975) of the centre of symmetry θ, together with an associated confidence interval. The Hodges–Lehmann estimate is defined as

$$\hat{\theta} = \text{median}\left\{\frac{x_i + x_j}{2}, 1 \leq i \leq j \leq n\right\}.$$

Let $m = (n(n+1))/2$ and let a_k, for $k = 1, 2, \dots, m$ denote the m ordered averages $\left(x_i + x_j\right)/2$ for $1 \leq i \leq j \leq n$. Then

if m is odd, $\hat{\theta} = a_k$ where $k = (m+1)/2$;
if m is even, $\hat{\theta} = (a_k + a_{k+1})/2$ where $k = m/2$.

This estimator arises from inverting the one-sample Wilcoxon signed-rank test statistic, $W(x - \theta_0)$, for testing the hypothesis that $\theta = \theta_0$. Effectively $W(x - \theta_0)$ is a monotonically decreasing step function of θ_0 with

$$\text{mean } (W) = \mu = \frac{n(n+1)}{4},$$

$$\text{var } (W) = \sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

The estimate $\hat{\theta}$ is the solution to the equation $W\left(x - \hat{\theta}\right) = \mu$; two methods are available for solving this equation. These methods avoid the computation of all the ordered averages a_k; this is because for large n both the storage requirements and the computation time would be excessive.

The first is an exact method based on a set partitioning procedure on the set of all ordered averages $\left(x_i + x_j\right)/2$ for $i \leq j$. This is based on the algorithm proposed by Monahan 1984.

The second is an iterative algorithm, based on the Illinois method which is a modification of the *regula falsi* method, see McKean and Ryan 1977. This algorithm has proved suitable for the function $W(x - \theta_0)$ which is asymptotically linear as a function of θ_0.

The confidence interval limits are also based on the inversion of the Wilcoxon test statistic.

Given a desired percentage for the confidence interval, $1 - \alpha$, expressed as a proportion between 0 and 1, initial estimates for the lower and upper confidence limits of the Wilcoxon statistic are found from

$$W_l = \mu - 0.5 + \left(\sigma\Phi^{-1}(\alpha/2)\right)$$

and

$$W_u = \mu + 0.5 + \left(\sigma\Phi^{-1}(1 - \alpha/2)\right),$$

where Φ^{-1} is the inverse cumulative Normal distribution function.

W_l and W_u are rounded to the nearest integer values. These estimates are then refined using an exact method if $n \leq 80$, and a Normal approximation otherwise, to find W_l and W_u satisfying

$$P(W \leq W_l) \leq \alpha/2$$
$$P(W \leq W_l + 1) > \alpha/2$$

and

$$P(W \geq W_u) \leq \alpha/2$$
$$P(W \geq W_u - 1) > \alpha/2.$$

Let $W_u = m - k$; then $\theta_l = a_{k+1}$. This is the largest value θ_l such that $W(x - \theta_l) = W_u$.

Let $W_l = k$; then $\theta_u = a_{m-k}$. This is the smallest value θ_u such that $W(x - \theta_u) = W_l$.

As in the case of $\hat{\theta}$, these equations may be solved using either the exact or the iterative methods to find the values θ_l and θ_u.

Then (θ_l, θ_u) is the confidence interval for θ. The confidence interval is thus defined by those values of θ_0 such that the null hypothesis, $\theta = \theta_0$, is not rejected by the Wilcoxon signed-rank test at the $(100 \times \alpha)\%$ level.

4 References

Lehmann E L 1975 *Nonparametrics: Statistical Methods Based on Ranks* Holden–Day

Marazzi A 1987 Subroutines for robust estimation of location and scale in ROBETH *Cah. Rech. Doc. IUMSP, No. 3 ROB 1* Institut Universitaire de Médecine Sociale et Préventive, Lausanne

McKean J W and Ryan T A 1977 Algorithm 516: An algorithm for obtaining confidence intervals and point estimates based on ranks in the two-sample location problem *ACM Trans. Math. Software* **10** 183–185

Monahan J F 1984 Algorithm 616: Fast computation of the Hodges–Lehman location estimator *ACM Trans. Math. Software* **10** 265–270

5 Parameters

5.1 Compulsory Input Parameters

1: **method – string**

Specifies the method to be used.

If **method** = 'E', the exact algorithm is used.

If **method** = 'A', the iterative algorithm is used.

Constraint: **method** = 'E' or 'A'.

2: **x(n) – double array**

The sample observations, x_i for $i = 1, 2, \ldots, n$.

3: **clevel – double scalar**

The confidence interval desired.

For example, for a 95% confidence interval set **clevel** = 0.95.

Constraint: 0.0 < **clevel** < 1.0.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the array **x**.

n, the sample size.

Constraint: $\mathbf{n} \geq 2$.

5.3 Input Parameters Omitted from the MATLAB Interface

wrk, iwrk

5.4 Output Parameters

1: **theta – double scalar**

The estimate of the location, $\hat{\theta}$.

2: **thetal – double scalar**

The estimate of the lower limit of the confidence interval, θ_l.

3: **thetau – double scalar**

The estimate of the upper limit of the confidence interval, θ_u.

4: **estcl – double scalar**

An estimate of the actual percentage confidence of the interval found, as a proportion between $(0.0, 1.0)$.

5: **wlower – double scalar**

The upper value of the Wilcoxon test statistic, W_u, corresponding to the lower limit of the confidence interval.

6: **wupper – double scalar**

The lower value of the Wilcoxon test statistic, W_l, corresponding to the upper limit of the confidence interval.

7: **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, **method** $\neq$ 'E' or 'A',
or **n** < 2,
or **clevel** ≤ 0.0,
or **clevel** ≥ 1.0.

ifail $= 2$

There is not enough information to compute a confidence interval since the whole sample consists of identical values.

ifail = 3

For at least one of the estimates $\hat{\theta}$, θ_l and θ_u, the underlying iterative algorithm (when **method** = 'A') failed to converge. This is an unlikely exit but the estimate should still be a reasonable approximation.

7 Accuracy

g07ea should produce results accurate to five significant figures in the width of the confidence interval; that is the error for any one of the three estimates should be less than $0.00001 \times$ (**thetau** − **thetal**).

8 Further Comments

The time taken increases with the sample size n.

9 Example

```
method = 'Exact';
x = [-0.23;
     0.35;
     -0.77;
     0.35;
     0.27;
     -0.72;
     0.08;
     -0.4;
     -0.76;
     0.45;
     0.73;
     0.74;
     0.83;
     -0.87;
     0.21;
     0.29;
     -0.91;
     -0.04;
     0.82;
     -0.38;
     -0.31;
     0.24;
     -0.47;
     -0.68;
     -0.77;
     -0.86;
     -0.59;
     0.73;
     0.39;
     -0.44;
     0.63;
     -0.22;
     -0.07000000000000001;
     -0.43;
     -0.21;
     -0.31;
     0.64;
     -1;
     -0.86;
     -0.73];
clevel = 0.95;
[theta, thetal, thetau, estcl, wlower, wupper, ifail] = g07ea(method, x,
clevel)
```

```
theta =
   -0.1300
```

```
thetal =
   -0.3300
thetau =
    0.0350
estcl =
    0.9514
wlower =
   556
wupper =
   264
ifail =
           0
```