NAG Toolbox for MATLAB

g07eb

1 Purpose

g07eb calculates a rank based (nonparametric) estimate and confidence interval for the difference in location between two independent populations.

2 Syntax

```
[theta, thetal, thetau, estcl, ulower, uupper, ifail] = g07eb(method, x,
y, clevel, 'n', n, 'm', m)
```

3 Description

Consider two random samples from two populations which have the same continuous distribution except for a shift in the location. Let the random sample, $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, have distribution $F(x)$ and the random sample, $y = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}$, have distribution $F(x - \theta)$.

g07eb finds a point estimate, $\hat{\theta}$, of the difference in location θ together with an associated confidence interval. The estimates are based on the ordered differences $y_j - x_i$. The estimate $\hat{\theta}$ is defined by

$$\hat{\theta} = \text{median}\left\{y_j - x_i, \qquad i = 1, 2, \ldots, n; j = 1, 2, \ldots, m\right\}.$$

Let d_k for $k = 1, 2, \ldots, nm$ denote the nm (ascendingly) ordered differences $y_j - x_i$ for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$. Then

if nm is odd, $\hat{\theta} = d_k$ where $k = (nm - 1)/2$;
if nm is even, $\hat{\theta} = (d_k + d_{k+1})/2$ where $k = nm/2$.

This estimator arises from inverting the two sample Mann–Whitney rank test statistic, $U(\theta_0)$, for testing the hypothesis that $\theta = \theta_0$. Thus $U(\theta_0)$ is the value of the Mann–Whitney U statistic for the two independent samples $\{(x_i + \theta_0)$, for $i = 1, 2, \ldots, n\}$ and $\left\{y_j, \text{ for } j = 1, 2, \ldots, m\right\}$. Effectively $U(\theta_0)$ is a monotonically increasing step function of θ_0 with

$$\text{mean}\,(U) = \mu = \frac{nm}{2},$$

$$\text{var}\,(U) = \sigma^2 \frac{nm(n + m + 1)}{12}.$$

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

The first is an exact method based on a set partitioning procedure on the set of all differences $y_j - x_i$ for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$. This is adapted from the algorithm proposed by Monahan 1984 for the computation of the Hodges–Lehmann estimator for a single population.

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 $U(\theta_0)$ which is asymptotically linear as a function of θ_0.

The confidence interval limits are also based on the inversion of the Mann–Whitney test statistic.

Given a desired percentage for the confidence interval, $1 - \alpha$, expressed as a proportion between 0.0 and 1.0 initial estimates of the upper and lower confidence limits for the Mann–Whitney U statistic are found;

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

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

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

U_l and U_u are rounded to the nearest integer values. These estimates are refined using an exact method, without taking ties into account, if $n + m \leq 40$ and $\max(n,m) \leq 30$ and a Normal approximation otherwise, to find U_l and U_u satisfying

$$\begin{array}{l} P(U \leq U_l) \leq \alpha/2 \\ P(U \leq U_l + 1) > \alpha/2 \end{array}$$

and

$$\begin{array}{l} P(U \geq U_u) \leq \alpha/2 \\ P(U \geq U_u - 1) > \alpha/2. \end{array}$$

The function $U(\theta_0)$ is a monotonically increasing step function. It is the number of times a score in the second sample, y_j, precedes a score in the first sample, $x_i + \theta$, where we only count a half if a score in the second sample actually equals a score in the first.

Let $U_l = k$; then $\theta_l = d_{k+1}$. This is the largest value θ_l such that $U(\theta_l) = U_l$.

Let $U_u = nm - k$; then $\theta_u = d_{nm-k}$. This is the smallest value θ_u such that $U(\theta_u) = U_u$.

As in the case of $\hat{\theta}$, these equations may be solved using either the exact or 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 Mann–Whitney two sample rank test at the $(100 \times \alpha)\%$ level.

4 References

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

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 observations of the first sample, x_i for $i = 1, 2, \ldots, n$.

3: **y(m) – double array**

The observations of the second sample, y_j for $j = 1, 2, \ldots, m$.

4: **clevel – double scalar**

The confidence interval required, $1 - \alpha$; e.g., 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 size of the first sample.

Constraint: **n** ≥ 1.

2: **m – int32 scalar**

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

m, the size of the second sample.

Constraint: **m** ≥ 1.

5.3 Input Parameters Omitted from the MATLAB Interface

wrk, iwrk

5.4 Output Parameters

1: **theta – double scalar**

The estimate of the difference in the location of the two populations, $\hat{\theta}$.

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: **ulower – double scalar**

The value of the Mann–Whitney U statistic corresponding to the lower confidence limit, U_l.

6: **uupper – double scalar**

The value of the Mann–Whitney U statistic corresponding to the upper confidence limit, U_u.

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** < 1,
or **m** < 1,
or **clevel** ≤ 0.0,
or **clevel** ≥ 1.0.

ifail $= 2$

Each sample consists of identical values. All estimates are set to the common difference between the samples.

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

g07eb should return 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 × (**thetau** – **thetal**).

8 Further Comments

The time taken increases with the sample sizes n and m.

9 Example

```
method = 'Approx';
x = [-0.582;
     0.157;
     -0.523;
     -0.769;
     2.338;
     1.664;
     -0.981;
     1.549;
     1.131;
     -0.46;
     -0.484;
     1.932;
     0.306;
     -0.602;
     -0.979;
     0.132;
     0.256;
     -0.094;
     1.065;
     -1.084;
     -0.969;
     -0.524;
     0.239;
     1.512;
     -0.782;
     -0.252;
     -1.163;
     1.376;
     1.674;
     0.831;
     1.478;
     -1.486;
     -0.8080000000000001;
     -0.429;
     -2.002;
     0.482;
     -1.584;
     -0.105;
     0.429;
     0.5679999999999999;
     0.944;
     2.558;
     -1.801;
     0.242;
     0.763;
     -0.461;
     -1.497;
     -1.353;
     0.301;
     1.941];
y = [1.995;
     0.007;
     0.997;
     1.089;
     2.004;
     0.171;
     0.294;
     2.448;
     0.214;
     0.773;
     2.96;
     0.025;
     0.638;
     0.9370000000000001;
     -0.5679999999999999;
     -0.711;
     0.931;
     2.601;
     1.121;
     -0.251;
     -0.05;
     1.341;
     2.282;
     0.745;
     1.633;
     0.944;
     2.37;
     0.293;
     0.895;
     0.9379999999999999;
     0.199;
     0.8120000000000001;
     1.253;
     0.59;
     1.522;
     -0.6850000000000001;
     1.259;
     0.571;
     1.579;
     0.5679999999999999;
     0.381;
     0.829;
     0.277;
     0.656;
     2.497;
     1.779;
     1.922;
     -0.174;
     2.132;
     2.793;
     0.102;
     1.569;
     1.267;
     0.49;
     0.077;
     1.366;
     0.056;
     0.605;
     0.628;
     1.65;
     0.104;
     2.194;
     2.869;
     -0.171;
     -0.598;
     2.134;
     0.917;
     0.63;
     0.209;
     1.328;
     0.368;
     0.756;
     2.645;
     1.161;
     0.347;
     0.92;
     1.256;
     -0.052;
     1.474;
     0.51;
     1.386;
     3.55;
     1.392;
     -0.358;
     1.938;
     1.727;
     -0.372;
     0.911;
     0.499;
     0.066;
     1.467;
     1.898;
     1.145;
     0.501;
     2.23;
     0.212;
     0.536;
     1.69;
     1.086;
     0.494];
clevel = 0.95;
[theta, thetal, thetau, estcl, ulower, uupper, ifail] = g07eb(method, x,
y, clevel)
```

```
theta =
    0.9505
thetal =
    0.5650
thetau =
    1.3050
estcl =
    0.9511
ulower =
        2007
uupper =
        2993
ifail =
          0
```