NAG Toolbox for MATLAB

g08ah

1 Purpose

g08ah performs the Mann–Whitney U test on two independent samples of possibly unequal size.

2 Syntax

```
[u, unor, p, ties, ranks, ifail] = g08ah(x, y, tail, 'n1', n1, 'n2', n2)
```

3 Description

The Mann–Whitney U test investigates the difference between two populations defined by the distribution functions $F(x)$ and $G(y)$ respectively. The data consist of two independent samples of size n_1 and n_2, denoted by $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$, taken from the two populations.

The hypothesis under test, H_0, often called the null hypothesis, is that the two distributions are the same, that is $F(x) = G(x)$, and this is to be tested against an alternative hypothesis H_1 which is

$H_1 : F(x) \neq G(y)$; or

$H_1 : F(x) < G(y)$, i.e., the x's tend to be greater than the y's; or

$H_1 : F(x) > G(y)$, i.e., the x's tend to be less than the y's,

using a two tailed, upper-tailed or lower-tailed probability respectively. You select the alternative hypothesis by choosing the appropriate tail probability to be computed (see the description of parameter **tail** in Section 5).

Note that when using this test to test for differences in the distributions one is primarily detecting differences in the location of the two distributions. That is to say, if we reject the null hypothesis H_0 in favour of the alternative hypothesis H_1: $F(x) > G(y)$ we have evidence to suggest that the location, of the distribution defined by $F(x)$, is less than the location, of the distribution defined by $G(y)$.

The Mann–Whitney U test differs from the Median test (see g08ac) in that the ranking of the individual scores within the pooled sample is taken into account, rather than simply the position of a score relative to the median of the pooled sample. It is therefore a more powerful test if score differences are meaningful.

The test procedure involves ranking the pooled sample, average ranks being used for ties. Let r_{1i} be the rank assigned to x_i, $i = 1, 2, \ldots, n_1$ and r_{2j} the rank assigned to y_j, $j = 1, 2, \ldots, n_2$. Then the test statistic U is defined as follows;

$$U = \sum_{i=1}^{n_1} r_{1i} - \frac{n_1(n_1 + 1)}{2}$$

U is also the number of times a score in the second sample precedes a score in the first sample (where we only count a half if a score in the second sample actually equals a score in the first sample).

g08ah returns:

(a) The test statistic U.

(b) The approximate Normal test statistic,

$$z = \frac{U - \text{mean}(U) \pm \frac{1}{2}}{\sqrt{\text{var}\,(U)}}$$

where

$$\text{mean}(U) = \frac{n_1 n_2}{2}$$

and

$$\text{var}(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \times TS$$

where

$$TS = \sum_{j=1}^{\tau} \frac{(t_j)(t_j - 1)(t_j + 1)}{12}$$

τ is the number of groups of ties in the sample and t_j is the number of ties in the jth group.

Note that if no ties are present the variance of U reduces to $\frac{n_1 n_2}{12}(n_1 + n_2 + 1)$.

(c) An indicator as to whether ties were present in the pooled sample or not.

(d) The tail probability, p, corresponding to U (adjusted to allow the complement to be used in an upper one-tailed or a two tailed test), depending on the choice of **tail**, i.e., the choice of alternative hypothesis, H_1. The tail probability returned is an approximation of p is based on an approximate Normal statistic corrected for continuity according to the tail specified. If n_1 and n_2 are not very large an exact probability may be desired. For the calculation of the exact probability see g08aj (no ties in the pooled sample) or g08ak (ties in the pooled sample).

The value of p can be used to perform a significance test on the null hypothesis H_0 against the alternative hypothesis H_1. Let α be the size of the significance test (that is, α is the probability of rejecting H_0 when H_0 is true). If $p < \alpha$ then the null hypothesis is rejected. Typically α might be 0.05 or 0.01.

4 References

Conover W J 1980 *Practical Nonparametric Statistics* Wiley

Neumann N 1988 Some procedures for calculating the distributions of elementary nonparametric teststatistics *Statistical Software Newsletter* **14 (3)** 120–126

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

5 Parameters

5.1 Compulsory Input Parameters

1: **x(n1) – double array**

The first vector of observations, $x_1, x_2, \ldots, x_{n_1}$.

2: **y(n2) – double array**

The second vector of observations. $y_1, y_2, \ldots, y_{n_2}$.

3: **tail – string**

Indicates the choice of tail probability, and hence the alternative hypothesis.

tail = 'T'

A two tailed probability is calculated and the alternative hypothesis is $H_1 : F(x) \neq G(y)$.

tail = 'U'

An upper-tailed probability is calculated and the alternative hypothesis $H_1 : F(x) < G(y)$, i.e., the x's tend to be greater than the y's.

tail = 'L'

A lower-tailed probability is calculated and the alternative hypothesis $H_1 : F(x) > G(y)$, i.e., the x's tend to be less than the y's.

Constraint: **tail** = 'T', 'U' or 'L'.

5.2 Optional Input Parameters

1: **n1 – int32 scalar**

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

the size of the first sample, n_1.

Constraint: **n1** ≥ 1.

2: **n2 – int32 scalar**

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

the size of the second sample, n_2.

Constraint: **n2** ≥ 1.

5.3 Input Parameters Omitted from the MATLAB Interface

wrk

5.4 Output Parameters

1: **u – double scalar**

The Mann–Whitney rank sum statistic, U.

2: **unor – double scalar**

The approximate Normal test statistic, z, as described in Section 3.

3: **p – double scalar**

The tail probability, p, as specified by the parameter **tail**.

4: **ties – logical scalar**

Indicates whether the pooled sample contained ties or not. This will be useful in checking which function to use should one wish to calculate an exact tail probability.

ties = **false**, no ties were present (use g08aj for an exact probability).

ties = **true**, ties were present (use g08ak for an exact probability).

5: **ranks(n1 + n2) – double array**

Contains the ranks of the pooled sample. The ranks of the first sample are contained in the first **n1** elements and those of the second sample are contained in the next **n2** elements.

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, **n1** < 1,
or **n2** < 1.

ifail $= 2$

On entry, **tail** $\neq$ 'T', 'U' or 'L'.

ifail $= 3$

The pooled sample values are all the same, that is the variance of $\mathbf{u} = 0.0$.

7 Accuracy

The approximate tail probability, p, returned by g08ah is a good approximation to the exact probability for cases where $\max(n_1, n_2) \geq 30$ and $(n_1 + n_2) \geq 40$. The relative error of the approximation should be less than 10%, for most cases falling in this range.

8 Further Comments

The time taken by g08ah increases with n_1 and n_2.

9 Example

```
x = [13;
     6;
     12;
     7;
     12;
     7;
     10;
     7;
     10;
     7;
     16;
     7;
     10;
     8;
     9;
     8];
y = [17;
     6;
     10;
     8;
     15;
     8;
     15;
     10;
     15;
     10;
     14;
     10;
     14;
     11;
```

```
      14;
      11;
      13;
      12;
      13;
      12;
      13;
      12;
      12];
tail = 'Lower-tail';
[u, unor, p, ties, ranks, ifail] = g08ah(x, y, tail)
```

```
u =
    86
unor =
   -2.8039
p =
    0.0025
ties =
     1
ranks =
   29.5000
    1.5000
   24.5000
    5.0000
   24.5000
    5.0000
   16.0000
    5.0000
   16.0000
    5.0000
   38.0000
    5.0000
   16.0000
    9.5000
   12.0000
    9.5000
   39.0000
    1.5000
   16.0000
    9.5000
   36.0000
    9.5000
   36.0000
   16.0000
   36.0000
   16.0000
   33.0000
   16.0000
   33.0000
   20.5000
   33.0000
   20.5000
   29.5000
   24.5000
   29.5000
   24.5000
   29.5000
   24.5000
   24.5000
ifail =
           0
```