Geofile Online

JANUARY 2005
485

ALAN MARRIOTT

Use of the Mann-Whitney U Test

Introduction to Significance Testing

Skills papers, especially at AS level, require the student to be familiar with statistics showing association or difference between two sets of data. Spearman's Rank Correlation Coefficient is a measure of association with which you may be familiar. This Geofile will help you with the Mann-Whitney U Test, which aims to show differences.

Significance testing is a mechanism for making decisions about the implications of sample data. We can never be entirely sure that a sample reflects the whole of the population, and significance testing quantifies the probability. In much geographical research the data available are sample data – eg data collected in a sampling process in the field – and we should always carry out an appropriate significance test before we draw conclusions from the data. In other cases we may use data derived from a full population survey and in this case we do not need to test for differences, changes, etc. If the full data show a difference, there *is* a difference. But, even in this situation, researchers sometimes use significance testing to test whether a difference is so small as to be of little interest or large enough to be of geographical importance.

The process of significance testing follows the same pattern whatever the test (see Box 1).

Box 1: Stages in Significance testing

Step 1: The Null Hypothesis
The assumption that there is no difference between two data sets, no change between two dates, or no association between two categorisations of data. The researcher is often looking to be able to reject the null hypothesis and make a more interesting statement about the research situation.

Step 2: The Significance Level
This is the chance of being wrong in rejecting the null hypothesis that the researcher is prepared to accept, usually 1% or 5% in geographical research.

Step 3: The Calculated Statistic
Obtained by applying the appropriate procedure, Mann-Whitney U in this case, to the data.

Step 4: The Critical Statistic
The critical statistic is determined by the sample sizes and the significance level. It is obtained by reference to published tables. Mann-Whitney U is different from other common statistical tests in that the critical value is the smallest, not the largest, value which is consistent with continued belief in the null hypothesis.

Step 5: The Decision
If the calculated Mann-Whitney U statistic is less than the critical value then the null hypothesis is rejected; if more the null hypothesis is accepted. In other words we decide that there is or is not a difference, has or has not been a change, etc.

Scales of Measurement

Numbers can be used in a variety of ways. If we are processing data on a computer we may quite arbitrarily use '0' for females and '1' for males. In this case the numerical values have no meaning at all. You can't sensibly work out averages, for example. Or numbers may be used to put items in an ordered list. The number scales defined below are in order of increasing rigour. Data obtained by measuring convey most information and are most valuable in research work.

The scales of measurement are:

1. Nominal – to do with naming, eg pass/fail, male/female. When working with nominal data you count the numbers in each category and calculate proportions/percentages, but not means, standard deviations etc.
2. Ordinal – relating to order, eg athletes in a race, candidates for a post. There are some statistical measures which use ordinal data and the appropriate measure of centrality to use is the median.
3. Interval – referring to the space between (i.e. measuring to establish differences between variables) eg ambient temperatures.
4. Ratio – interval data where the relative differences are established (ie where there is a zero). An example is the time it takes a floating object to travel along the length of an underground stream.

Temperature data are not like this, however, since there is not a genuine zero. Neither 0°C nor 0°F are real zeros, and they are different, of course. So we can't say that 20°C is twice as hot as 10°C, although we can say that a time elapse of 2 minutes 12 seconds is twice 1 minute 6 seconds. Interval and ratio data give more information than ordinal data and the full set of powerful statistical techniques can be used with them.

Choice of Significance Test

The most sensitive and powerful significance tests, like t tests, are very widely used. But they do depend on certain conditions being met. These conditions about the distribution of data in the populations from which the samples are taken are, first, that the data are normally distributed, and second, if we are testing two samples against one another, that the two populations have the same standard deviation. However, often we work with populations which do not meet these conditions. In these cases we use alternative tests, distribution-free tests or non-parametric tests, which are easy to use and are increasingly popular in geographical research. In Box 2 some of the most common research situations are listed and appropriate significance tests indicated.

The Mann-Whitney U Test

The Mann-Whitney U test is one of the most frequently used distribution-free significance tests. It is widely used in geographical research, although it was first developed in the late 1940s by statisticians undertaking industrial research. The test assesses whether the degree of overlap between the two observed distributions is more than would be expected by chance, on the null hypothesis that the two samples are drawn from a single population or from populations with the same overall level of rankings.

The test involves the calculation of a statistic, called U. For small samples, the distribution is tabulated in many books, but for samples above about 20 (some statisticians would allow lower sample sizes), the normal distribution is a good approximation.

Box 2: Choice of significance test

1. Testing for differences in the general level of data, eg differences between means or medians

	Population normally distributed	**Population distribution unknown or not normal**
Level of Measurement	Ratio or interval data	Ordinal data - or ratio or interval data reduced to ordinal
Testing a single sample against a known population value	t test	Wilcoxon
Testing two samples against one another - data paired	t test	Wilcoxon
Testing two samples against one another - data not paired	t test	Wilcoxon Rank Sum & Mann-Whitney U Test
Testing several samples against one another	F test (ANOVA)	Kruskal-Wallis Test

2. Testing for differences in the frequency with which different outcomes occur

	Population distribution not relevant
Level of Measurement	Nominal data – or ordinal, interval or ratio data reduced to nominal
All situations	Chi-squared Test

Box 3: The first distribution-free statistical test?

The greater number of male births compared with female births is now well known to students of demography. But what is commonplace now was not understood before the eighteenth century. A paper published in 1710 used 82 years of birth records for the City of London. In every one of those years the number of boys born exceeded the number of girls. If the probabilities of male and female births were the same, we would expect more or less equal numbers of years when male births outnumbered female births and vice versa. If equal probabilities did exist, the probability of 82 male majorities in a sample of 82 years would be incredibly low (in fact a decimal number with 25 zeros before the first significant digit).

Although the author did not express his working in this way, he was, in practice, carrying out a distribution-free significance test. The implied null hypothesis was that there were equal numbers of years in which the births of each sex exceeded the other. The test used involved the use of signs (+ and -) and assumed no knowledge of the underlying distribution. And the null hypothesis was clearly rejected.

Note that a test based on the normal distribution (or better still the t distribution) might have been even more effective since the method used neglected a lot of information. But this was a century before the discovery of the normal distribution and two centuries before the discovery of the t distribution.

Worked Example 1: Geology and Attractiveness of the National Parks of England and Wales

In Table 1 the National Parks of England and Wales are listed in order of their attractiveness as suggested by a senior officer of a Geographical Association branch. (We should note that (1) the data were generated by a subjective on-the-spot process, but they are rank data and so illustrate the use of this test; (2) the data can be regarded as a sample of the ranking that all the senior officers of GA branches could give.) We wish to use these rankings to determine whether there is any difference in the assessment of attractiveness according to the rock formations of the National Parks, which are largely igneous and metamorphic or largely sedimentary. There are three largely igneous and metamorphic National Parks, the Lake District, Dartmoor and Snowdonia, which will comprise one sample; the other eight the second sample.

Step 1: Null hypothesis

Null hypothesis: There is no statistically significant evidence that there is a difference in the rankings of igneous/metamorphic and sedimentary National Parks.

Tip: The null hypothesis is a rather uninteresting statement which the researcher usually wishes to disprove. The non-committal way in which null hypotheses are worded often (but not always) means that we do not learn much about geography unless we have enough information to reject them.

Step 2: Significance level

At 5% significance level

Tip: The normal advice about significance levels in geographical research is to choose 1% or 5%. However, there seem to be few 1% tables published so, in practice, you may have to use the 5% level.

Step 3. Calculating Mann-Whitney U

The calculations are very straightforward, requiring simple arithmetic only.

Step 3.1: Rank the combined sample values from lowest to highest (see Table 2). Tied ranks are given the mean of the tied values.

Step 3.2: Sum the ranks for each sample separately to obtain R_S and R_L, the total of the ranks for the smaller sample and the larger sample.

Step 3.3: Calculate U_S using:

$$\begin{aligned} U_S &= n_S n_L + (n_S(n_S + 1))/2 - R_S \\ &= 3 \text{ x } 8 + (3(3 + 1))/2 - 16 \\ &= 24 + 6 - 16 \\ &= 14 \end{aligned}$$

Step 3.4: Calculate U_L using:

$$\begin{aligned} U_L &= n_S n_L - U_S \\ &= 3 \text{ x } 8 - 14 \\ &= 24 - 14 \\ &= 10 \end{aligned}$$

or:

$$\begin{aligned} U_L &= n_S n_L + (n_L(n_L + 1))/2 - R_L \\ &= 3 \text{ x } 8 + (8(8 + 1))/2 - 50 \\ &= 24 + 36 - 50 \\ &= 10 \end{aligned}$$

Table 1: The National Parks of England and Wales ranked in order of attractiveness

1. Lake District
2. Pembrokeshire Coast
3. Peak District
4. North York Moors
5. Yorkshire Dales
6. Dartmoor
7. Exmoor
8. Brecon Beacons
9. Snowdonia
10. New Forest
11. Northumberland

Step 3.5: Define U_{calc} as the smaller of U_S and U_L.
Hence:
$U_{calc} = 10$

Step 4: Critical value of U
This is obtained from published tables, for $n_S = 3$ and $n_L = 8$ at the 5% significance level.
Hence:
$U_{critical,\, n=3 \text{ and } 8,\, 0.05} = 2$

Step 5: State decision
The rule here is that calculated values less than the critical value show a significant difference between the two samples.
In this case
$10 > 2$
$U_{calc} > U_{critical}$

Accept the null hypothesis
There is no evidence that igneous/metamorphic National Parks are ranked differently to sedimentary National Parks in terms of their attractiveness.

Tip: In the Mann-Whitney U test the calculated statistic has to be *less* than the critical statistic for us to reject the null hypothesis. This is very unusual; in most cases calculated statistic has to be more than the critical value. So think carefully! In Mann-Whitney U the critical statistic is not a hurdle to be cleared, but a limbo stick to be danced under.

Worked Example 2: Corruption and Economic Growth

In recent years the World Bank, the International Monetary Fund and bilateral aid donors including the UK have increasingly linked the rate of economic development with the quality of governance. 'Governance' includes the related themes of efficiency and fairness in public administration and decision making. Good government is built on objectives shared by the people of the states concerned (most would regard economic growth as a good thing, for example), it implements policies which promote those shared objectives, and it applies those policies impartially and objectively. A criticism that has been made of many developing countries is that the governments' objectives are sometimes not those of the people (for example, an unspoken objective might be to enrich the ruling class, not the bulk of the population) and that even where there are policies, administrative procedures and laws intended to promote development, these are by-passed by ineffective enforcement and, worse, corrupt practices.

To test whether there is a relationship between corruption and economic growth, we will examine some data for growth rates in two samples of states: 12 which are classified as having a high level of corruption, and 13 which have low level of corruption (see Table 3).

The growth data are measured on the ratio scale but there is no reason to expect them to be normally distributed. So we will rank the values and use the Mann-Whitney U to test whether there is a difference between low and high corruption countries.

Notes on the data

1. The per capita income data are taken from the *World Bank Atlas* of 1997 and 2003. The 1995 data are Gross National Product per capita expressed in Purchasing Power Parity terms, ie the data are adjusted to allow for different costs of living in the countries concerned. The 2001 data are Gross National Income per capita, also given is PPP terms. (GNP and GNI are identical, but the World Bank's preferred terminology has changed.)
2. The classification of the countries is taken from the work of Transparency International, an NGO based in the USA, quoted by Gray and Kaufman (1998). In total 45 states were classified – the remaining states were intermediate between the two groups given. Vietnam was one of the high corruption states but was excluded because there were no income data published in the *World Bank Atlas*.

Steps 1 and 2: Null hypothesis and significance level
There is no statistically significant evidence that the general level of income growth differs between states with low and high levels of corruption (at 5% significance level)

Table 2: Ranking of sample values; Column A = igneous/metamorphic National Parks, Column B = sedimentary National Parks

National Park	A	B
Lake District	1	
Pembrokeshire Coast		2
Peak District		3
North York Moors		4
Yorkshire Dales		5
Dartmoor	6	
Exmoor		7
Brecon Beacons		8
Snowdonia	9	
New Forest		10
Northumberland		11
Sum of Ranks	16	50

Step 3: Calculation of the U statistic

$U_S = n_S n_L + (n_S(n_S + 1))/2 - R_S$
$= 12 \text{ x } 13 + (12(12 + 1))/2 - 153$
$= 156 + 78 – 153$
$= 81$
$U_L = n_S n_L – U_S$
$= 12 \text{ x } 13 – 81$
$= 156 – 81$
$= 75$
or: $U_L = n_S n_L + (n_L(n_L + 1))/2 - R_L$
$= 12 \text{ x } 13 + (13(13 + 1))/2 – 172$
$= 156 + 91 – 172$
$= 75$
Hence: $\mathbf{U_{calc} = 75}$

Box 4: Definitions of U

U_{calc} = the actual value of the Mann-Whitney U statistic for the samples being considered (the smaller of U_S and U_L).
$U_{critical}$ = the value determined by the choice of significance level.
U_S = the calculated value for the smaller sample (some books use U_1 for this).
U_L = the calculated value for the larger sample (sometimes U_2)
Note: The reason for choosing the smaller sample is for arithmetical simplicity, the result is not affected by reversing the two.

Box 5: The Formula for Mann-Whitney U

$U_S = n_S n_L + (n_S(n_S + 1))/2 - R_S$
$U_L = n_S n_L – U_S$
or:
$U_L = n_S n_L + (n_L(n_L + 1))/2 - R_L$
(where R_S and R_L are the sums of the ranks of the smaller and larger samples)
U_{calc} is the smaller of U_S and U_L.
Normal distribution approximation:
$\mu = n_1 n_2/2$
$\sigma = \surd[n_1 n_2(n_1 + n_2 + 1)/12]$

Table 3: Income and corruption data for selected countries

	Level of Corruption	GNP per capita (PPP $) 1995	GNI per capita (PPP $) 2001	Percentage Change 1995–2001	High Corruption Rank	Low Corruption Rank
Argentina	High	8,310	10,980	32.1	17	
Australia	Low	18,940	24,630	30.0		15
Brazil	High	5,400	7,070	30.9	16	
Canada	Low	21,130	26,530	25.6		10
China	High	2,920	3,950	35.3	19	
Colombia	High	6,130	6,790	10.8	6	
Denmark	Low	21,230	28,490	34.2		18
Finland	Low	17,760	24,030	35.3		20
Germany	Low	20,070	25,240	25.8		11
India	High	1,400	2,820	101.4	25	
Indonesia	High	3,800	2,830	-25.5	2	
Ireland	Low	15,680	27,170	73.3		24
Israel	Low	16,490	19,630	19.0		8
Mexico	High	6,400	8,240	28.8	14	
Netherlands	Low	19,950	27,390	37.3		21
New Zealand	Low	16,360	18,250	11.6		7
Philippines	High	2,850	4,070	42.8	22	
Russia	High	4,480	6,880	53.6	23	
Singapore	Low	22,770	22,850	0.4		4
Sweden	Low	18,540	23,800	28.4		13
Switzerland	Low	25,860	30,970	19.8		9
Thailand	High	7,540	6,230	-17.4	3	
Turkey	High	5,580	5,830	4.5	5	
United Kingdom	Low	19,260	24,340	26.4		12
Venezuela	High	7,900	5,590	-29.2	1	
Totals					153	172

Step 4: Determination of critical value
$U_{\text{critical, n = 12 and 13, 0.05}} = 41$

Step 5: Decision
Since $U_{calc} > U_{critical}$
Accept the null hypothesis. There is no evidence of a difference.

Worked Example 3: Corruption and Economic Growth (Use of Normal Distribution)

When the two samples are larger the sampling distribution is approximately normal. We will illustrate by re-working the corruption and economic growth example.

Steps 1 and 2
These are the same.
Null hypothesis: There is no statistically significant evidence that the general level of income growth differs between states with low and high levels of corruption (at 5% significance level)

Step 3: Calculation of the z statistic
The mean and standard deviation of the sampling distribution are calculated as follows:

Mean $= n_S n_L/2$
$= 12(13)/2$
$= 156/2$
$= 78$

Standard deviation
$= \sqrt{[n_S n_L (n_S + n_L + 1)/12]}$
$= \sqrt{[12 \text{ x } 13(12 + 13 + 1)/12]}$
$= \sqrt{[3900/12]} = \sqrt{338} = 18.38$

U_{calc} (which is 81, see above) is then used as follows:
$z = (U_{calc} – \text{Mean of U})/\text{Standard deviation of U}$
$= (81 – 78)/18.38$
$z_{calc} = 0.163$

Step 4: Critical value of z
From tables of the normal distribution
$z_{\text{critical, 0.05}} = 1.96$

Step 5: Decision
Since $z_{calc} < z_{critical}$
Accept the null hypothesis. There is no evidence of a difference.

Tip: Since the calculations are a lot easier with the direct method for Mann-Whitney U, you are advised to avoid the normal approximation as far as possible. However, there are not many tables for sample sizes over 20, so if at least one of the two sample sizes is 20 or more you will probably not have a choice. Some statisticians suggest the use of the normal approximation with sample sizes as low as 10, or even 8, suggesting that the accuracy of the result is not seriously affected.

References

Gray, C.W. and Kaufman, D. (1998) 'Corruption and Development' *Finance and Development,* March.

International Bank of Reconstruction and Development (1997 and 2003) *World Bank Atlas,* Washington DC.

Focus Questions

1. (a) Give examples of situations in which you may generate (i) ordinal data and (ii) interval or ratio data. If possible draw your examples from your own fieldwork.
(b) Illustrate the need for an objective approach for deciding on the significance of your results when you are working with sample data. Again, draw your examples from your own fieldwork.

2. Do the data in Worked Example 1 suggest a different view of the attractiveness of English and Welsh (Pembrokeshire Coast, Brecon Beacons, and Snowdonia) National Parks? (Check: $U_{calc} = 11$.)

3. The hypothesis examined in Worked Example 2 was probably inappropriate in that there may well be so strong a link between corruption and level of development (rather than the rate of development, which we tested) that other effects are swamped. Rank the countries by GNI per capita in 2001 and carry out a Mann-Whitney U test. (Check: $U_{calc} = 0$.)