Tutorials in Quantitative Methods for Psychology
2008, vol. 4(1), p. 13-20.

The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution

Nadim Nachar
Université de Montréal

It is often difficult, particularly when conducting research in psychology, to have access to large normally distributed samples. Fortunately, there are statistical tests to compare two independent groups that do not require large normally distributed samples. The Mann-Whitney U is one of these tests. In the following work, a summary of this test is presented. The explanation of the logic underlying this test and its application are presented. Moreover, the forces and weaknesses of the Mann-Whitney U are mentioned. One major limit of the Mann-Whitney U is that the type I error or alpha *(α)* is amplified in a situation of heteroscedasticity.

It is generally recognized that psychological studies often involve small samples. For example, researchers in clinical psychology often have to deal with small samples that generally include less than 15 participants (Kazdin 2003; Shapiro & Shapiro, 1983; Kraemer, 1981; Kazdin, 1986). Although the researchers aim at collecting large normally distributed samples, they rarely have the appropriate amount of resources (time and money) to recruit a sufficient number of participants. It is thus useful, particularly in psychology, to consider tests that have few constraints and allow experimenters to test their hypotheses on small and poorly distributed samples.

A lot of studies do not provide very good tests for their hypotheses because their samples have too few participants (for a review of the reviews, see Sedlmeier & Gigerenzer, 1989). Even tough small samples can be methodologically questionable (e.g. generalization is difficult); they can be useful to infer conclusions on the population if the adequate statistical test is applied.

One can imagine a situation where a scientist has two groups of subjects but has only very few participants in each group (less than eight participants). Thus, this researcher cannot affirm that his two groups come from a normal distribution because they include too few participants (Mann and Whitney, 1947). In addition to this statistical "constraint", the data of the research conducted by this experimenter is of continuous or ordinal type. This implies that his measurements can be lacking in precision. In such a case, this researcher cannot refer to the parametric test of mean using the Student's t-distribution because it is impossible to check that the two samples are normally distributed. How can one react in such a situation? Initially, a statistical test of non-parametric type imposes itself for this researcher (a non-parametric test is necessary when the distribution is asymmetrical). Non-parametric tests differ from parametric test in that the model structure is not specified a priori but determined from the data. The term *nonparametric* is not meant to imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance. Therefore, nonparametric tests are also called *distribution free*. The Mann-Whitney U test can be used to answer the questions of the researcher concerning the difference between his groups. This test has the great advantage of possibly being used for small samples of subjects (five to 20 participants). It can also be used when the measured variables are of ordinal type and were recorded with an arbitrary and not a very precise scale.

In the field of behavioural sciences, the Mann-Whitney U test is one of the most commonly used non-parametric

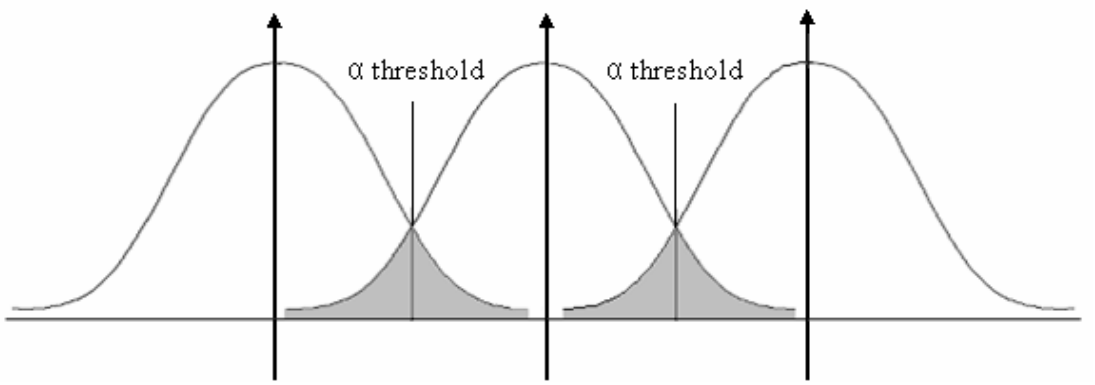

Figure 1. Normal distributions illustrating one-tailed and two-tailed tests

statistical tests (Kasuya, 2001). This test was independently worked out by Mann and Whitney (1947) and Wilcoxon (1945). This method is thus often called the Wilcoxon-Mann-Whitney test or the Wilcoxon sum of ranks test.

In the following text, a brief summary of the Mann and Whitney method will be presented. The underlying logic of this test, an example of its application as well as the use of SPSS for its calculation will be presented. Lastly, some forces and limits of the test will be reported.

1. The Mann-Whitney U Test

1.1. Hypotheses of the Test

The Mann-Whitney U test null hypothesis *(H_0)* stipulates that the two groups come from the same population. In other terms, it stipulates that the two independent groups are homogeneous and have the same distribution. The two variables corresponding to the two groups, represented by two continuous cumulative distributions, are then called stochastically equal.

If a two-sided or two-tailed test is required, the alternative hypothesis *(H_1)* against which the null hypothesis is tested stipulates that the first group data distribution differs from the second group data distribution. In this case, the null hypothesis is rejected for values of the test statistic falling into either tail of its sampling distribution (see Figure 1 for a visual illustration). On the other hand, if a one-sided or one-tailed test is required, the alternative hypothesis suggests that the variable of one group is stochastically larger than the other group, according to the test direction (positive or negative). Here, the null hypothesis is rejected only for values of the test statistic falling into one specified tail of its sampling distribution (see Figure 1 for a visual illustration).

Table 1. Numbers of social phobia's symptoms after the therapy

Behavioral therapy (*B*)	Combined therapy (*C*)
3	1
3	1
4	2
4	2
7	5
7	5
7	5

The data of the table are fictitious.

In more specific terms, let one imagine two independent groups that have to be compared. Each group contains a number n of observations. The Mann-Whitney test is based on the comparison of each observation from the first group with each observation from the second group. According to this, the data must be sorted in ascending order. The data from each group are then individually compared together. The highest number of possible paired comparisons is thus: $(n_x n_y)$, where n_x is the number of observations in the first group and n_y the number of observations in the second. If the two groups come from the same population, as stipulated by the null hypothesis, each datum of the first group will have an equal chance of being larger or smaller than each datum of the second group, that is to say a probability p of one half (1/2). In technical terms,

H_0: $p(x_i > y_j) = 1/2$ and

H_1: $p(x_i > y_j) \neq 1/2$

(two-tailed test) where x_i is an observation of the first sample and y_j is an observation of the second.

The null hypothesis is rejected if one group is significantly larger than the other group, without specifying the direction of this difference.

In a one-tailed application of the test, the null hypothesis remains the same. However, a change is brought to the alternative hypothesis by specifying the direction of the comparison. This relation can be expressed mathematically,

Table 2. Numbers of social phobia's symptoms after the therapy and their ranks

Numbers of symptoms	1	1	2	2	3	3	4	4	5	5	5	7	7	7
Behavioral therapy (*b*) / Combined therapy (*c*)	*c*	*c*	*c*	*c*	*b*	*b*	*b*	*b*	*c*	*c*	*c*	*b*	*b*	*b*
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The data of the table are fictitious.

H_0: $p(x_i > y_j) = 1/2$ and H_1: $p(x_i > y_j) > 1/2$.

This alternative hypothesis implies that the quantity of elements, or the dependent variable measurements, of the first group are significantly larger than those of the second. Note that the groups can be interchanged, in which case the alternative hypothesis corresponds to:

H_1: $p(x_i > y_j) < 1/2$.

The hypotheses previously quoted can also be in terms of medians. The null hypothesis states that the medians of the two respective samples are not different. As for the alternative hypothesis, it affirms that one median is larger than the other or quite simply that the two medians differ. In a more explicit way, the hypothesis respectively corresponds to:

H_0: $\theta_x = \theta_y$, H_1: $\theta_x < \theta_y$ or $\theta_x > \theta_y$ (one-tailed test)

H_0: $\theta_x = \theta_y$, H_1: $\theta_x \neq \theta_y$ (two-tailed test)

where θ_x corresponds to the median of the first group and θ_y corresponds to the median of the second group.

Therefore if the null hypothesis is not rejected, it means that the median of each group of observations are similar. On the contrary, if the two medians differ, the null hypothesis is rejected. The two groups are then considered as coming from two different populations.

1.2. Assumptions of the Test

In order to verify the hypotheses, the sample must meet certain conditions. These conditions can be easily respected. They are of three types:

(a) The two investigated groups must be randomly drawn from the target population. The concept of random implies the absence of measurement and sampling errors (Robert et al., 1988). Note that an error of these last types can be involved but must remain small.

(b) Each measurement or observation must correspond to a different participant. In statistical terms, there is independence within groups and mutual independence between groups.

(c) The data measurement scale is of ordinal or continuous type. The observations values are then of ordinal, relative or absolute scale type.

1.3. The Test

The Mann-Whitney U test initially implies the calculation of a U statistic for each group. These statistics have a known distribution under the null hypothesis identified by Mann and Whitney (1947) (see Tables 3 to 8).

Mathematically, the Mann-Whitney U statistics are defined by the following, for each group:

$$U_x = n_x n_y + \left(\left(n_x(n_x+1)\right)/2\right) - R_x \quad (1)$$

$$U_y = n_x n_y + \left(\left(n_y(n_y+1)\right)/2\right) - R_y \quad (2)$$

where n_x is the number of observations or participants in the first group, n_y is the number of observations or participants in the second group, R_x is the sum of the ranks assigned to the first group and R_y is the sum of the ranks assigned to the second group.

In other words, both U equations can be understood as the number of times observations in one sample precede or

Table 3. Probability of Obtaining a U not Larger than that Tabulated in Comparing Two Samples when $n_x = 3$

	n_y		
U	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.200
3		.600	.350
4			.500
5			.650

Table 4. Probability of Obtaining a U not Larger than that Tabulated in Comparing Two Samples when $n_x = 4$

	n_y			
U	1	2	3	4
0	.200	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

Table 5. Probability of Obtaining a U not Larger than that Tabulated in Comparing Two Samples $n_x = 5$

	n_y				
U	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

Table 6. Probability of Obtaining a U not Larger than that Tabulated in Comparing Two Samples when $n_x = 6$

	n_y					
U	1	2	3	4	5	6
0	.143	.036	.012	.005	.002	.001
1	.286	.071	.024	.010	.004	.002
2	.428	.143	.048	.019	.009	.004
3	.571	.214	.083	.033	.015	.008
4		.321	.131	.057	.026	.013
5		.429	.190	.086	.041	.021
6		.571	.274	.129	.063	.032
7			.357	.176	.089	.047
8			.452	.238	.123	.066
9			.548	.305	.165	.090
10				.381	.214	.120
11				.457	.268	.155
12				.545	.331	.197
13					.396	.242
14					.465	.294
15					.535	.350
16						.409
17						.469
18						.531

follow observations in the other sample when all the scores from one group are placed in ascending order (see the Procedure and Application section for further information). In this respect, note that the order in which the data is arranged is unique when the measurement scale is of continuous type. Following the assumption of continuity, two observations cannot take the same value.

Following the calculation of the U statistics and the determination of an appropriate statistical threshold (α), the null hypothesis can be rejected or not. In other words, there is rejection of H_0 if, by consulting the Mann and Whitney tables, the p corresponding to the $min(U_x, U_y)$ (the smallest of U both calculated) is smaller than the p or the predetermined α threshold. In technical terms,

Reject H_0 if p of $min(U_x, U_y) < \alpha$ *threshold*.

1.4. Normal approximation

If the numbers of observations n_x and n_y are larger than eight, a normal approximation, as shown by Mann and Whitney (1947), can be used, that is to say:

$$\mu_U = (nxny)/2 = (U_x + U_y)/2 \text{ and}$$
$$\sigma_U = \sqrt{((nxny)(N+1))/12}$$

where $N = (nx + ny)$, μ_U corresponds to the average of the U distribution and σ_U corresponds to its standard deviation.

If each group includes more than eight observations, the sample's distribution gradually approaches a normal distribution. If a normal approximation has to be used, the corresponding equation becomes:

$$z = (U - (nxny/2))/\sigma U$$

and the test statistic becomes, in absolute values:

$$|z| = |U_x + U_y|/\sigma U \ .$$

To test the difference between $U_{x \text{ or } y}$ and μ_U, the reader can refer to the z-table. If the absolute value of the calculated z is larger or equal to the tabulated z value, the null hypothesis is rejected.

Reject H_0 if $|calculated\ z| \geq |z\ tabulated|$.

1.5. Ties (equalities)

Following the Mann-Whitney U test assumption of continuity, the data's arrangement must be unique. This postulate implies that it is impossible that two values are exactly equal (chances are one out of infinity). However, it is often possible to observe equal measurements in behavioural sciences because the measurements are rarely very precise. In the case of equalities, it is necessary to calculate both U by allocating half of the tied ranks (ties) to the first group's values and the other half to the second group's values. It is as if one gave to each observation, the average rank if no equality had existed. Note that when ties

occur within a group, this type of equality does not need to be considered in the calculation presented here. Indeed, it is the equalities between the two groups that deserve attention. In short, in the event of ties, assign the rank of half of the observations to the first group and the other half to the second.

In a situation of ties between the groups, the normal approximation must be used with an adjustment to the standard deviation. The standard deviation or the square root of the variance becomes:

$$\sigma U = \sqrt{\left((nxny)/(N(N-1))\right)\left(\left(N^3-N\right)/12\right)-\sum_{j=1}^{g}\left(\left(tj^3-tj\right)/12\right)}$$

where $N=(n_x+n_y)$, g = number of ties and t_j = number of equal ranks in the second group.

2. Procedure and Application

Let take a fictitious example of application of the U test.

Table 7. Probability of Obtaining a U not Larger than that Tabulated in Comparing Two Samples when $n_x = 7$

	n_y						
U	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

An experimenter read that there is an antibiotic often tested, well documented, and known to help information storage in memory. This experimenter also knows through scientific reports and guidelines that the behavioral therapy has an established efficacy for the treatment of the social phobia (APA, 1998; INSERM, 2004; BPSCORE, 2001). In addition, he knows that the behavioral therapy requires the learning of new behaviours which implies information storage.

The number of symptoms of social phobia after two types of therapy was investigated. Two groups of individuals with social phobia were compared. The first group received the behavioral therapy; the second group received the behavioral therapy combined with the antibiotic. After each therapy, both groups showed a decreased in the number of symptoms of social phobia. The number of these symptoms was measured and a test was run to decide whether the combined therapy had more effect on the symptoms than the behavioral therapy alone.

In other terms, the experimenter wishes to compare two random variables having continuous cumulative distribution functions. He wishes to test the hypothesis that his variables are stochastically equal (their distributions are similar) against the alternative that C is stochastically smaller than B. C corresponds to the numbers of symptoms under investigation in the combined therapy group, B corresponds to the numbers of symptoms in the behavioral therapy group.

Unfortunately, the number of subjects with social phobia is small.

Moreover, nothing indicates that the symptomatology is normally distributed amongst the individuals with social phobia. Hence, the Mann and Whitney U test is the only legitimate test. Table 1 shows the results of the experiment.

First, organize each group data in ascending order irrespective of group membership. Be aware that a value of *-20* is ordered, on an increasing scale, before a value of *-10*. See Table 2 for a visual illustration.

Both U statistics can be computed using the equations (1) and (2).

Note that the sum of ranks of the two groups is always $Rx + Ry$:

$$(R_x + R_y) = 1+2+3+...+N = \left(N(N+1)\right)/2$$

where $N=(n_x+ny)$

$$Rx + Ry = (nx+ny)(nx+ny+1)/2 = N(N+1)/2 \qquad (3)$$

In this way, one can deduce, starting from the equations (1) and (2), that:

$$R_x = n_x n_y + \left(\left(n_x(n_x+1)\right)/2\right) - U_x$$
$$R_y = n_x n_y + \left(\left(n_y(n_y+1)\right)/2\right) - U_y$$

Inserting these two preceding equations in equation (3):

$$(R_x + R_y) = n_x n_y + \left(\left(n_x(n_x+1)\right)/2\right) - U_x + n_x n_y + \left(\left(n_y(n_y+1)\right)/2\right) - U_y$$
$$= \left((n_x+n_y)(n_x+n_y+1)\right)/2$$

so that

$$n_x n_y + \left(\left(n_x(n_x+1)\right)/2\right) + n_x n_y + \left(\left(n_y(n_y+1)\right)/2\right) - \left((n_x+n_y)(n_x+n_y+1)\right)/2$$
$$= U_x + U_y$$

Table 8. Probability of Obtaining a U not Larger than that Tabulated in Comparing Two Samples when $n_x = 8$

	n_y								
U	1	2	3	4	5	6	7	8	normal
0	.111	.022	.006	.002	.001	.000	.000	.000	.001
1	.222	.044	.012	.004	.002	.001	.000	.000	.001
2	.333	.089	.024	.008	.003	.001	.001	.000	.001
3	.444	.133	.042	.014	.005	.002	.001	.001	.001
4	.556	.200	.067	.024	.009	.004	.002	.001	.002
5		.267	.097	.036	.015	.006	.003	.001	.003
6		.356	.139	.055	.023	.010	.005	.002	.004
7		.444	.188	.077	.033	.015	.007	.003	.005
8		.556	.248	.107	.047	.021	.010	.005	.007
9			.315	.141	.064	.030	.014	.007	.009
10			.387	.184	.085	.041	.020	.010	.012
11			.461	.230	.111	.054	.027	.014	.016
12			.539	.285	.142	.071	.036	.019	.020
13				.341	.177	.091	.047	.025	.026
14				.404	.217	.114	.060	.032	.033
15				.467	.262	.141	.076	.041	.041
16				.533	.311	.172	.095	.052	.052
17					.362	.207	.116	.065	.064
18					.416	.245	.140	.080	.078
19					.472	.286	.168	.097	.094
20					.528	.331	.198	.117	.113
21						.377	.232	.139	.135
22						.426	.268	.164	.159
23						.475	.306	.191	.185
24						.525	.347	.221	.215
25							.389	.253	.247
26							.433	.287	.282
27							.478	.323	.318
28							.522	.360	.356
29								.399	.396
30								.439	.437
31								.480	.481
32								.520	

then:

$$U_x + U_y = n_x n_y$$

The sum of U_x and U_y is thus equal to the product of the two samples sizes. Consequently, once one value of U is obtained using the equations (1) or (2), the value of the other U is found by subtracting the value of the first U from the product of the two samples sizes. Thus:

$$U_x = (n_x n_y) - U_y \text{ or } U_y = (n_x n_y) - U_x$$

This last equation can save an enormous amount of time.

Second, the researcher must calculate the U statistic corresponding to each group using equations (1) or (2):

$$U_B = n_B n_C + \left(\left(n_B(n_B+1)\right)/2\right) - R_B$$
$$= 7\times 7 + \left(7(7+1)\right)/2 - (5+6+7+8+12+13+14)$$
$$= 49 + 28 - 65 = 12$$

so that $U_C = 7\times 7 - 12 = 37$. Alternatively:

$$U_C = n_B n_C + \left(\left(n_C(n_C+1)\right)/2\right) - R_C$$
$$= 7\times 7 + \left(7(7+1)\right)/2 - (1+2+3+4+9+10+11)$$
$$= 49 + 28 - 40 = 37$$

where n_B is the number of symptoms of social phobia after the behavioral therapy, n_C is the number of symptoms of social phobia after the combined therapy, R_B is the sum of the ranks assigned to the behavioral therapy group and R_C is the sum of the ranks assigned to the combined therapy group.

Based on this method, the experimenter can formulate the null hypothesis differently: U_B does not differ significantly from U_C.

Third, compute the global U statistic in this way: $\min(U_x, U_y)$ (choose the smallest value of both U statistics calculated). With the Mann and Whitney tables (1947), the probability of obtaining a U value that is not larger than the one calculated above can be obtained. To find this probability in the Mann and Whitney tables, the following information is required: the value of $\min(U_x, U_y)$, n_x and n_y. If the probability of obtaining such a U is smaller than the predetermined alpha (α) threshold, the null hypothesis is rejected. If it is a one-tailed test, this value found in the Mann and Whitney tables correspond to the probability *p* value (probability of rejecting H_0 when this one is "true") which will be compared with the predetermined alpha (α) threshold of statistical significance. On the other hand, if it is a two-tailed test, it is necessary to double this probability to obtain the one that will be compared with the predetermined alpha *(α)* threshold of statistical significance.

Referring to Table 7, the smallest U is in this case 12 and correspond to a *p* of 0.064 (see also SPSS section). Thus, this *p* is not smaller than a predetermined *p* of, for example, 0.05. The researcher does not reject H_0 and concludes that the two groups are not significantly different.

3. Computing the Mann-Whithney U test using SPSS

First of all, one needs to enter the data in *SPSS*, not forgetting the golden rule which stipulates that each participant's observation must occupy a line. The numbers of the groups are generally 1 and 2, except whenever it is more practical to use other numbers.

Following the entry of the data, open a new syntax window and enter the following syntax.

```
NPAR TESTS
  /M-W= name of the dependent variable column BY
name of the independent variable column (1 2)
  /STATISTICS= DESCRIPTIVES QUARTILES
  /MISSING ANALYSIS.  or  /MISSING LISTWISE.
```

The last line of the preceding syntax corresponds to two options that manage the missing values. These options are useful when more than one statistical test is specified in the syntax table. The first option is /MISSING ANALYSIS and supports that each test is separately evaluated for the missing values. On the other hand, with the option /MISSING LISTWISE, each empty box or missing value, for any variable, is excluded from all analyses. The option that one will choose depends on the other tests that one needs to apply. If there is only the Mann-Whitney statistical test that has to be carried out, the missing values will be managed in the same manner, does not matter which of the two options is selected.

Following the syntax execution, the results appear in tables in the *Output* window. Initially, descriptive data like the group averages, their standard deviation, the minimal and maximal values, the quartiles and the number of participants in each group appear. Thereafter, the test results appear in two distinct tables. In the first table, between the values of the *Ranks*, the *Mean Rank* and the *Sum of Ranks* given, the *N* corresponds to the number of observations or participants. In addition, in the second one, the tests results appear. *SPSS* automatically provides us the Mann-Whitney U, the Wilcoxon W and the Z results. This computer program also returns the asymptotic significance or the level of significance based on the normal distribution of the statistical test: *Asymp. Sig. (2-tailed):*. In a general way, a value lower than the statistical threshold is considered significant and the alternative hypothesis is accepted.

The asymptotic significance is based on the assumption that the data sample is large. If the data sample is small or badly distributed, the asymptotic significance is not in general a good indication of the significance. In this case, the level of significance based on the exact distribution of a statistical test or *Exact Sig. [2*(1-tailed Sig.)]* corresponds to the statistic of decision. Consequently, one should use this value when the sample is small, sparse, contains many ties, is badly balanced or does not seem to be normally distributed. SPSS thus provides the exact value of *p* (*Exact Sig. [2*(1-tailed Sig.)]*) and the value of *p* based on a normal approximation (*Asymp. Sig. (2-tailed)*). If a normal distribution is adequate to the studied case, the two values should be roughly or exactly equivalent. Note that *Asymp. Sig. (2-tailed):* and *Sig. [2*(1-tailed Sig.)]:* represent two level of significance for a two-tailed test. If one uses a one-tailed test, these two levels must be divided by two. Lastly, the mention *Not corrected for ties:* imply that the test did not correct the result appearing in the table for the ties or equalities.

According to the example previously presented, the researcher will consider the *Exact Sig. [2*(1-tailed Sig.)]:* . This done and because his test application is of one-tailed type, he will divide this level of significance based on the exact distribution by two to obtain the level of significance that will be compared to his predetermined statistical threshold *(α).* In the example previously presented, the *p* is 0.064 and not smaller than the predetermined statistical threshold of 0.05.

4. Discussion

Like any statistical test, the Mann-Whitney U has forces and weaknesses. In terms of forces, like any non-parametric test, the Mann-Whitney U does not depend on assumptions on the distribution (i.e. one does not need to postulate the data distribution of the target population). One can also use it when the conditions of normality neither are met nor realisable by transformations. Moreover, one can use it when his sample is small and the data are semi-quantitative or at least ordinal. In short, few constraints apply to this test.

The Mann-Whitney U test is also one of the most powerful non-parametric tests (Landers, 1981), where the statistical power corresponds to the probability of rejecting a false null hypothesis. This test has thus good probabilities of providing statistically significant results when the alternative hypothesis applies to the measured reality. Even if it is used on average-size samples (between 10 and 20 observations) or with data that satisfy the constraints of the t-test, the Mann-Whitney has approximately 95% of the Student's t-test statistical power (Landers). By comparison with the t-test, the Mann-Whitney U is less at risk to give a wrongfully significant result when there is presence of one or two extreme values in the sample under investigation (Siegel and Castellan, 1988).

Despite this, the Mann and Whitney test (1947) has its limits. With the Monte Carlo methods, methods that calculate a numerical value by using random or probabilistic processes, it was shown that the t-test is most of the time more powerful than the U-test. Indeed, this fact remains whatever the amplitude of the differences between the averages of the populations under investigation and even if

the distributions of these populations do not meet the criteria of normality (Zimmerman, 1985). On the other hand, very little statistical power is lost if the Mann-Whitney U test is used instead of the t-test and this, under statistically controlled conditions (Gibbons and Chakraborti, 1991).

In addition, the Mann-Whitney U test is, in exceptional circumstances, more powerful than the t-test. Indeed, it is more powerful in the detection of a difference on the extent of the possible differences between populations' averages than the t-test when a small manpower is associated with a small variance (Zimmerman, 1987). On the other hand, when the sample size is similar or when the smallest manpower has the greatest variance, the t-test is more powerful on all the extent of the possible differences (Zimmerman).

Lastly, the Monte Carlo methods showed that the Mann-Whitney U test can give wrongfully significant results, that is to say the erroneous acceptance of the alternative hypothesis (Robert & Casella, 2004). This type of results is at risk to be obtained whenever one's samples are drawn from two populations with a same average but with different variances. In this type of situations, it is largely more reliable to use the t-test which gives a possibility for the samples to come from distributions with different variances. The alpha (α) error or of type I is to reject H_0 whereas this one is true. This error is thus amplified when Mann-Whitney U is applied in a situation of heteroscedasticity or distinct variances. In addition, some solutions exist to this major problem (see Kasuya, 2001).

In short, the Mann-Whitney U statistical test is an excellent alternative to parametric tests like the t-test, when the assumptions of these last ones cannot be respected. With a statistical power similar to the t-test, the Mann-Whitney U is, by excellence, the test of replacement. However, as one understood, it is more reliable to use the t-test if its postulates can be met.

References

APA - American Psychological Association (1998). Special section: Empirically supported psychological therapies. *Journal of Consulting and Clinical Psychology, 66(1).*

BPSCORE - British Psychological Society Centre for Outcomes Research and Effectiveness (2001). *Treatment Choice in Psychological Therapies and Counselling: Evidence Based Clinical Practice Guideline.* Royaume-Uni: Department of Health.

Gibbons, J.D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's t, and alternate t tests for means of normal distributions. *Journal of Experimental Education,* 59(3), 258-267.

INSERM - Institut national de la santé et de la recherche médicale (2004). *Psychothérapies: trios approches évaluées*. Paris: Édition INSERM.

Kasuya, E. (2001). Mann-Whitney U test when variances are unequal. *Animal Behavior, 61*, 1247-1249.

Kazdin, A. E. (1986). Comparative outcome studies in psychotherapy: Methodological issues and strategies. *Journal of Consulting and Clinical Psychology, 54,* 95-105.

Kazdin, A.E. (2003). *Methodological Issues and Strategies in Clinical Research (3rd edition).* Washington, D.C.: American Psychological Association.

Kraemer, H. C. (1981). Coping strategies in psychiatric clinical research. *Journal of Consulting and Clinical Psychology, 49*, 309-319.

Landers, J. (1981). *Quantification in History, Topic 4 : Hypothesis Testing II-Differing Central Tendency*. Oxford : All Souls College.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of 2 random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50-60.

Robert, C.P., & Casella, G. (2004). *Monte Carlo Statistical Methods, second edition.* New York: Springer-Verlag.

Robert, M. et al. (1988). *Fondements et étapes de la recherche scientifique en psychologie.* Saint-Hyacinthe : Edisem et Paris : Maloine.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 42-53.

Siegel, S., & Castellan, N.J.Jr. (1988) *Nonparametric statistics for the behavioral sciences, second edition*. États-Unis : McGraw-Hill book company.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80-83.

Zimmerman, D.W. (1987). Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. *Journal of Experimental Education*, *55*, 171-174.

Zimmerman, D.W. (1985). Power functions of the t test and Mann-Whitney U test under violation of parametric assumptions. *Perceptual and Motor Skills*, *61*, 467-470.

Manuscript received 29 September 2006
Manuscript accepted 1 May 2007