Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 8

5-2016

A Two and More Independent Factor Form of the Wilcoxon-Mann-Witney Test, Extendable to Other Permutation-Based Tests

Christopher H. Holland
King's College London, christopher.c.holland@kcl.ac.uk

Tom Holland
Triveritas Ltd., st.win@ntlworld.com

A Two and More Independent Factor Form of the Wilcoxon-Mann-Witney Test, Extendable to Other Permutation-Based Tests

Cover Page Footnote

This work has its origin in the Fellowship thesis for the Royal College of Veterinary Surgeons of one of us. We would like to thank the thesis supervisor Peter Lee for his invariably prompt and helpful criticisms of this work. We would like to thank Niclas Sjogren (latterly of AstraZeneca Statistics Department) for the fruitful discussions of the ideas from which this work came. The Fellowship work was initially supported by AstraZeneca.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 123-140

ISSN 1538 − 9472

A Two and More Independent Factor Form of the Wilcoxon-Mann-Witney Test, Extendable to Other Permutation-Based Tests

Christopher H. Holland
King's College London
London, United Kingdom

Tom Holland
Triveritas Ltd.
Brampton, Cumbria, United Kingdom

The Wilcoxon-Mann-Witney test is extended to account for a second independent factor. The new test statistic's probability mass function and normal approximation are derived. Critical-values for balanced, unbalanced, and large sample designs are given. The immediate extension of this method to a wide range of non-parametric tests is explained.

Keywords: Two factor nonparametric test, Wilcoxon-Mann-Witney test, nuisance factors, rank, order, permutation

Introduction

The Wilcoxon-Mann-Witney test (WMWt) is a widely used technique for data analysis in which a natural ordering is possible. For example, it may be possible to order or rank the subjective degree of inflammation or pain; yet objective measurement of the inflammation or pain may be impossible. The WMWt is simple, robust and powerful; it has a minimum asymptotic relative efficiency (ARE or Pitman efficiency) of 95% compared to Students' *t* test, the most efficient test possible under ideal conditions (Conover, 1999). With small samples, heteroscedastic, and non-normal data the WMWt can have much greater power than Students *t* test (Blair, Higgins, & Smitely, 1980). The WMWt is widely taught and used because of these properties – it also has great intuitive appeal to experimenters' common sense.

However, it is common that experiments are run in which an extraneous factor (such as sex of the subject or strain of animal) is included which is not of direct experimental interest and may confound the result. When this occurs, it is unreasonable to ignore the 'nuisance' factor in the analysis of ranked/ordered data

Mr. Christopher Holland is a graduate medical student in the clinical portion of his training. Email him at: christopher.c.holland@kcl.ac.uk.

on the effect of the treatment. For example, in assessing the effect of a treatment on pain or inflammation by ranking of pain or inflammation in both sexes, it would be unreasonable to simply ignore the sex of the subjects. This is because the sex of an experimental subject substantially influences both the nature and degree of an inflammatory response and the perception of pain in man and animals (reviewed by Berkley, Zalcman, & Simon, 2006), and can thus confound any result.

A practical method of extending the WMWt is proposed so that a nuisance factor which divides the sample into two distinct sets for which direct comparison of subjects between sets is precluded (such as subject sex) can be accounted for in the detection of a significant effect of the explanatory variable upon the dependent variable. As such, this nuisance factor can be excluded from influencing the conclusions drawn. The proposed extension unifies an overall analysis of otherwise subsetted data and so offers increased statistical power and avoids conflicting conclusions between data subsets. It also avoids the ambiguous situation where a series of tests on small subsets of the data by the simple WMWt may give different results (e.g. the males show a significant treatment effect which is not shown by the females). Measuring interactions between factors (i.e. detecting if there is a difference in the degree of response of the males compared to the females) is not possible in this simple formulation given here. However, a substantially more complex extension of our techniques makes this possible (theoretical approach outlined in Holland, 2011).

Heuristic Development of the Statistical Model

Consider the one-tailed Mann-Witney formulation of the U statistic (Sprent & Smeeton, 2001). This test is used to compare two groups of subjects, made distinct by some factor (such as in an animal study, a treated group and a control group) for which the subjects in the two groups can be ordered in some feature (for example an experimental parameter). For the remainder of the article we refer to the two groups as a control group (denoted c) and a treated group (t) for convenience. The purpose is to deduce whether there is a statistically significant difference in this feature between the two groups. To obtain a U statistic from the data, one takes each element of control group and counts the number of samples in the treated group (t) that show less of the feature. The final U value is obtained by summing this count for each element of the control group.

Hence *least* tcctc *most* $U = 4$
cttcc $U = 4$
while tctcc $U = 5$

To assess statistical significance, we wish to compare this value with that expected assuming the null hypothesis. Under the null hypothesis that there is no difference between the two groups, every ordering of samples is equally likely. Thus the key step in assigning significance to a particular U-value obtained from an experiment is to identify the probability mass function under the null hypothesis by quantifying the number of different permutations of the given numbers of treated and controls that give rise to each possible value of U. A result is significant if the U-value falls beyond a certain point in the extreme tail of this probability mass function.

The usual assumption underpinning the WMWt applies to this work, namely: the samples are independent of each other. So this is a test simply for a difference in location or, equivalently, a difference in the mean (if it exists) and median between the two groups.

The proposed test procedure accounts for the scenario in which the subjects in the sample are additionally divided by a second factor into two sets (e.g. male and female) where comparison of individuals between these two sets is precluded, effectively making it impossible to calculate the U-statistic of the whole sample. Here, the first two U statistics are calculated by first considering the two sets separately and calculating the U statistic of each with respect to the feature of interest (labeled U_1 and U_2). For example, if sex is this second factor, a U_1 statistic for the males is produced and a U_2 statistic for the females is produced in the usual way, considering the two sets as completely separate; as such, no comparison of males to females or females to males is made.

However, if the analysis ends here, then two different statistics are produced despite both being the result of an identical experimental design, and the data has effectively been subsetted. This causes a number of difficulties. For instance, due to the reduction in the group sizes from the original, each would have reduced statistical power for a given overall sample size. There is also the potential for the two results to have similar but conflicting implications (for example just managing to achieve significance in one set, while just failing to achieve significance in the other), making it unclear as to the overall conclusion.

To avoid such problems, it is proposed that instead of ending the analysis there, the two separately obtained U-values (U_1 and U_2) are added together to form a *combined U-value* U_C. This approach has the very appealing property that each comparison of any individual pair contributes exactly the same amount of information to the overall test statistic as any other comparison, and so any difference of sample size requires no correction or weighting of the contributing U_1 or U_2 values to U_C.

To analyze this result further, the U_C is then compared to the distribution of U_C statistics expected under the null hypothesis to infer the statistical significance of any overall difference between the two groups, similarly to the one-factor case.

Establishing the number of distinct permutations giving rise to each possible U_C value, and thus establishing the probability mass function of the U_C statistic under the null hypothesis, is the real substance of this article. We then use this to tabulate critical values for all balanced two-factor study designs up to a group size of 10 for one- and two-tailed designs. For unbalanced and larger group sizes we give a computer algorithm that gives an exact probability mass function and a simple normal approximation method.

Methodology

Revisiting the Analysis for a Single Factor *U* Statistic

First, to aid in understanding the two set case, we reexamine the well-established case of a single set of directly comparable subjects divided into two groups: a treated group with m subjects and a control group with n subjects. Consider the scenario in which, for an appropriate feature, the ordering of the members of the two groups is found and the U-value is calculated. The aim is to establish whether this U-value represents a statistically significant difference between the two groups by comparing it to what is expected under the null hypothesis: that there is no overall difference in the feature between the two groups.

Under the null hypothesis, this definition implies that, for any given subject, its position in the ordering is not influenced at all by which group it is from. Hence, if the null hypothesis is true, each of the orderings of the subjects is equally likely. Furthermore, we may ignore the internal orderings of both the treated and control groups and so just consider the distinct permutations of m indistinguishable t's and n indistinguishable c's as each such permutation is equally likely. For example, in case $m = 2$, $n = 3$, there would be an equal probability of an experiment giving rise to the ordering ‘tctcc’ as ‘ctctc’ (or any other distinct permutation of two t's and three c's, for that matter). This is because such a permutation represents $m!n!$ different subject orderings. As this number is constant for all permutations and each ordering is equally likely, each of the permutations is each equally likely under the null hypothesis.

Each of these distinct permutations has an associated U-value. The next step is therefore to establish the number of distinct permutations that give rise to each U-value. The function $f_1(m, n \mid r)$ is introduced for this purpose. This represents, for treated and control groups of size *m* and *n* respectively, the number of permutations that give a *U*-value of *r*. For example, in an experiment involving 2 treated and 3 controls,

$$\mathrm{f}_1\left(2,3\,|\,r\right)=1,1,2,2,2,1,1$$

for r = 0, 1, 2, 3, 4, 5, 6, respectively. So, for instance, by looking at the 5th element in the sequence, we can deduce that there are two permutations giving rise to a *U*-value of 4. These are tcctc and cttcc.

As previously established under the null hypothesis, the distinct permutations are each equally likely; this shows that, in such conditions, it would be twice as likely to obtain a *U*-value of 3 as a *U*-value of 1 (for example). As such, for fixed *m* and *n*, the value of $\mathrm{f}_1(m, n\,|\,r)$ represents the relative probability of achieving the different values of *U* (represented by *r*) under the null hypothesis. Therefore, if this function is normalized for fixed *m* and *n*, we obtain the probability mass function for *U* under the null hypothesis:

$$\mathrm{p}_{1:m,n}\left(r\right)=\frac{\mathrm{f}_1\left(m,n\,|\,r\right)}{{}^{m+n}C_m}\ . \tag{1}$$

This uses the property

$$\sum_{i=1}^{mn}\mathrm{f}_1\left(m,n\,|\,i\right)={}^{m+n}C_m\ . \tag{2}$$

Hence as it effectively yields the distribution of *U*-values expected under the null hypothesis, $\mathrm{f}_1(m, n\,|\,r)$ is the key function in establishing whether an experimentally obtained *U*-value represents a statistically significant difference between the two groups or not. No closed form of $\mathrm{f}_1(m, n\,|\,r)$ exists, but it satisfies recursion relations which allow its calculation using a computer (some such recursive properties are reviewed in Di Bucchianico (1999)).

Due to the importance of $\mathrm{f}_1(m, n\,|\,r)$, we discuss some of its properties. First, we note that there are no permutations yielding a *U*-value which is negative or in excess of *mn*. As such, $\mathrm{f}_1(m, n\,|\,r) = 0$ unless $0\leq r \leq mn$. Secondly, the sum over all possible values of $\mathrm{f}_1(m, n\,|\,r)$ is given by (2).

Finally, for fixed *m*, *n*, $\mathrm{f}_1(m, n\,|\,r)$ is symmetric in *r* about the central value, *mn*/2, so that

$$\mathrm{f}_1\left(m,n\,|\,r\right)=\mathrm{f}_1\left(m,n\,|\,mn-r\right)\ . \tag{3}$$

This also means that both the mean and median value of *U* under the null hypothesis are *mn*/2.

In order to find $\mathrm{f}_1(m, n\,|\,r)$ for given *m*, *n*, use the method suggested by Theorem 2.6 in Di Bucchianico (1999), which we derived without the use of the correspondence with the restricted partition function. This utilizes the relations

$$\begin{aligned}
&\mathrm{f}_1\left(m,n\,|\,r\right)=0 \quad \text{for } r<0, r>mn, m<0, n<0\\
&\mathrm{f}_1\left(m,n\,|\,0\right)=1 \quad \text{for } m\geq 0, n\geq 0\\
&\mathrm{f}_1\left(m,n\,|\,r\right)=\mathrm{f}_1\left(m-1,n\,|\,r-n\right)+\mathrm{f}_1\left(m,n-1\,|\,r\right) \quad \text{for } 1\leq r\leq mn, m\geq 1, n\geq 1
\end{aligned}$$

Using these formulae, it is possible to specify $\mathrm{f}_1(m, n\,|\,r)$ for any finite values of *m*, *n*, or *r*. The Matlab 7.8.0 program which we used to find $\mathrm{f}_1(m, n\,|\,r)$ for given *m* and *n* and for the possible values of *r* is given in Appendix 1. This was able to calculate $\mathrm{f}_1(20, 20\,|\,r)$ for all possible *r* practically instantaneously, and $\mathrm{f}_1(100, 100\,|\,r)$ within 100 seconds using a home laptop computer.

Having obtained the probability mass function under the null hypothesis, it is a simple matter to analyze the statistical significance of a given result. For example, it can be used to calculate the *P*-value once an experimental *U*-value is obtained or a specified confidence interval for the null value.

However, as the values of *m* and *n* become very large, it becomes unfeasibly onerous even for a computer to calculate $\mathrm{f}_1(m, n\,|\,r)$ exactly. In that case, approximations may be of greater practical use. For sets where each group size exceeds 20, the normal approximation is deemed sufficiently accurate for most usual cases. This involves approximating the probability mass function $p_{1;m,n}(r)$ by a normal distribution, with mean *mn*/2 and standard deviation

$$\sigma_U=\sqrt{\frac{mn\left(m+n+1\right)}{12}}\ . \tag{4}$$

Defining the Frequency Distribution of the U_C Statistic

As outlined above, the aim is to extend the analysis to the case where there are two sets which differ in some factor other than the experimental parameter, such that an element of one set cannot be reasonably compared to an element of the other set. A *combined U-value* U_C is formed by finding a *U*-statistic for each set separately and then adding the two values. As in the single set case, the goal is to determine the statistical significance of an experimentally-obtained U_C-value.

By identical reasoning as in the single set case, if the null hypothesis holds, then each combination of distinct permutations of the treated and control subjects in each set is equally likely. As before, this observation leads to the construction of a function that enumerates the number of permutation combinations that give rise to a given U_C-value for specified group sizes.

Start by establishing a function: $\mathrm{F}_2(m, n; p, q \mid r; s)$. This function gives the number of permutation combinations which simultaneously achieve: a U_1-value of r in set one (with treated group size m and control group size n) and a U_2-value of s in set two (with treated group size p and control group size q). Due to the independence of the two sets, this is the product of the number of permutations in set 1 giving a U-value of r and the number of permutations in set 2 giving a U-value of s. Alternatively, using the notation of the previous section:

$$\mathrm{F}_2\left(m,n;p,q\,|\,r,s\right)=\mathrm{f}_1\left(m,n\,|\,r\right)*\mathrm{f}_1\left(p,q\,|\,s\right)\ . \tag{5}$$

This can be represented, for given m, n, p, and q, as a matrix with the row denoted by r and the column by s. An example of this is given in Table 1, where it is detailed for $\mathrm{F}_2(2, 2; 2, 3 \mid r; s)$.

From $\mathrm{F}_2(m, n; p, q \mid r; s)$, the function can be created giving the number of permutation combinations giving rise to a U_C-value of k: $\mathrm{f}_2(m, n; p, q \mid k)$. This is the equivalent of $\mathrm{f}_1(m, n \mid r)$ for the two set case. It is constructed by adding together all elements of $\mathrm{F}_2(m, n; p, q \mid r; s)$ with the specified values of m, n, p, q where the sum of r and s is equal to k, as can be seen in (6) below:

Table 1. The two-dimensional U frequency array for an $\mathrm{F}_2(2, 2; 2, 3 \mid r; s)$ study design, illustrating how the sum of each trailing diagonal gives the total number of ways of achieving each value of U_C (U_C = 3 cells are identified explicitly)

	U_2 value (*s*)	**0**	**1**	**2**	**3**	**4**	**5**	**6**
U_1 value (*r*)	Individual frequency	1	1	2	2	2	1	1
0	1	1	1	2	2*	2	1	1
1	1	1	1	2*	2	2	1	1
2	2	2	2*	4	4	4	2	2
3	1	1*	1	2	2	2	1	1
4	1	1	1	2	2	2	1	1

Note: *U_C = 3

$$\mathrm{f}_2\left(m,n;p,q\,|\,k\right)=\sum_{i+j=k}\mathrm{F}_2\left(m,n;p,q\,|\,i;j\right)\ . \tag{6}$$

In the matrix representation, this is simply adding the entries along the appropriate trailing diagonal, which is shown in Table 1 for F$_2$(2, 2; 2, 3 | *r*; *s*) for the U_C value of *k* = 3. By summing the indicated numbers, we find that f$_2$(2, 2; 2, 3 | 3) = 7. To illustrate the point, these seven permutation combinations for males and females respectively may be identified as:

> tctc and ccctt (*U*-values 3 & 0 respectively); tcct and cctct (2 & 1); cttc and cctct (2 &1); ctct and ccttc (1 & 2); ctct and ctcct (1 & 2); cctt and ctctc (0 & 3); cctt and tccct (0&3)

By carrying out similar sums along the other trailing diagonals,

$$\mathrm{f}_2\left(2,2;2,3\,|\,k\right)=1,2,5,7,10,10,10,7,5,2,1$$

for *k* = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, respectively.

Several properties of f$_2$(*m*, *n*; *p*, *q* | *k*) follow directly from the properties of f$_1$(*m*, *n* | *r*). For instance, the function is non-zero only for $0 \le k \le (mn + pq)$, and the overall number of permutations regardless of U_C-value (i.e. the sum of f$_2$(*m*, *n*; *p*, *q* | *k*) over all *k* for given group sizes) is

$$\sum_{i=0}^{mn+pq}\mathrm{f}_2\left(m,n;p,q\,|\,i\right)={}^{m+n}C_m * {}^{p+q}C_p\ . \tag{7}$$

Somewhat less trivially, f$_2$(*m*, *n*; *p*, *q* | *k*) is also symmetric in *k* around the central value (*mn* + *pq*)/2. This is seen as follows:

$$\begin{aligned}\mathrm{f}_2\left(m,n;p,q\,|\,(mn+pq)-k\right)&=\sum_{i+j=(mn+pq)-k}\mathrm{f}_1\left(m,n\,|\,i\right)\mathrm{f}_1\left(p,q\,|\,j\right)\\&=\sum_{(mn-i)+(pq-j)=k}\mathrm{f}_1\left(m,n\,|\,mn-i\right)\mathrm{f}_1\left(p,q\,|\,pq-j\right)\\&=\sum_{i'+j'=k}\mathrm{f}_1\left(m,n\,|\,i'\right)\mathrm{f}_1\left(p,q\,|\,j'\right)\\&=\mathrm{f}_2\left(m,n;p,q\,|\,k\right)\end{aligned}$$

On line 3, dummy variables i and j were replaced by dummy variables $i' = pq - i$ and $j' = pq - j$, respectively. The mean and median of $f_2(m, n; p, q \mid k)$ are therefore both $(mn + pq)/2$. Finally, as for $f_1(m, n \mid r)$, we can use $f_2(m, n; p, q \mid k)$ to construct the probability mass function for U_C:

$$p_{2;m,n;p,q}(k) = \frac{f_2(m,n;p,q \mid k)}{{}^{m+n}C_m \, {}^{p+q}C_p} \ . \tag{8}$$

These properties allow a U_C-value to be analyzed using the same approach as the U-value in the single set case, just using $f_2(m, n; p, q \mid k)$ instead of $f_1(m, n \mid r)$ to, for example, calculate the P-value of a given U_C or a confidence interval under the null hypothesis. This is discussed further in the section below. The recursive Matlab program we used to find $f_2(m, n; p, q \mid k)$ for given m, n, p, and q, and for the possible values of k is given in Appendix 2.

Again, it is possible to use a normal approximation to $p_{2;m,n;p,q}(k)$ for large group sizes, ideally all groups in excess of 10. This is accomplished simply by the distribution resulting from the addition of normally distributed random variables of the two separate normal approximations of the two sets. This results in a normal distribution with mean $(mn + pq)/2$ and standard deviation

$$\sigma_{U_C} = \sqrt{\frac{mn(m+n+1) + pq(p+q+1)}{12}} \ . \tag{9}$$

Table 2. The values that U_C must Equal or Exceed to achieve significance at the given confidence level for one-tailed and two-tailed tests

	One-Tailed		Two-Tailed	
Group Size	**95%**	**99%**	**95%**	**99%**
1	-	-	-	-
2	8	-	-	-
3	15	17	2, 16	0, 18
4	25	28	5, 27	3, 29
5	37	41	11, 39	7, 43
6	52	57	18, 54	13, 59
7	68	75	26, 72	20, 78
8	87	96	37, 91	29, 99
9	108	119	49, 113	39, 123
10	132	144	62, 138	52, 148

Results

Using the Exact Distribution of U_C

As discussed above, obtaining the probability mass function under the null hypothesis allows us to establish confidence intervals for the null hypothesis. This gives us a range of U_C-values for which the null hypothesis can be rejected for a specified statistical significance (α). This gives us critical values of U_C, which are the highest (or lowest) value of U_C such that the null hypothesis cannot be rejected. The most commonly used values of α are 0.05 and 0.01, which give a 95% and 99% confidence interval, respectively, for either one- or two-tailed tests. As such, the critical values of U_C for all experiments with uniform group sizes (i.e. $m = n = p = q$) up to 10 are given in Table 2 for these confidence intervals. A dash is used for circumstances in which the test can never yield a statistically significant result.

To calculate the critical values for other cases, such as unbalanced designs or for data in which group sizes are larger than 10, a program such as that in Appendix 2 can be used to find the $f_2(m, n; p, q \mid k)$ function from which the probability mass function can be obtained. Any confidence interval under the null hypothesis can then be constructed and, similarly, an exact P-value can be calculated.

Using the Asymptotic Distribution of U_C

Where m, n, p, and q are greater than 10, the normal approximation can be used (derived from Campbell (1974), including continuity correction). The distribution of z follows the Normal distribution and the one- and two-tailed critical values can be taken directly from standard tables:

$$z = \frac{\left(U_C \pm \frac{1}{2} - \frac{(mn + pq)}{2}\right)}{\sqrt{\frac{1}{12}\left(mn(m+n+1) + \left(pq(p+q+1)\right)\right)} \tag{10}$$

The values of m, n, p, and q at which this approximation becomes workably accurate for the 95% confidence level, both one- and two-tailed, are investigated using tables that give 3 significant figures for z. The exact p value is computed for all combinations of group sizes up to 10 (i.e. 4 groups each of 10 or less samples) for all U_C values possible, and the U_C value which just exceeds the 95% confidence limit is found. Hence the next U_C value is obtained closer to the mean than this, the highest non-significant U_C value. Then, compare this lowest significant U_C highest non-significant U_C pair to the value given by the normal approximation form of the test. The normal approximation test is very close to the exact value unless one or more groups have a group size of one. Except for this extreme condition, the one-tailed test normal approximation never gives a significant result as non-significant. However, in 31 of 1035 one-tailed cases, the normal approximation gives as significant a U_C that the exact test gives as just non-significant. The group sizes that give a false positive are given in Table 3; all are the result of rounding errors, and the normal approximation gives complete concordance with the exact test working to 7 figure accuracy. In the two-tailed test, none of the highest non-significant U_C values give a significant finding, but the approximate methods fails to detect significance in 91 of 1035 two-tailed test. The group sizes that give a false negative are given in Table 4.

We think this remarkable accuracy of the normal approximation is the result of the action of the Central Limit Theorem when combining two mass functions that are themselves fairly close to normal. Practically, it means that so long as ${}^{m+n}C_m * {}^{p+q}C_p > 20$ (one-tailed) and 41 (two-tailed), then (given that the 95% confidence limit is an arbitrary cut-off point) the normal approximation is a practical method for all study designs that do not use group sizes of 1.

Table 3. One-tailed 95% confidence limit; these are the groups sizes that are just not significant at the exact 95% confidence limit, but in the normal approximations are just significant (*z* = ±1.65 to 3 significant figures for all them)

2,	4:	8,	9	3,	4:	7,	10
2,	5:	4,	9	3,	5:	7,	9
2,	4:	9,	10	4,	10:	7,	10
2,	4:	9,	9	4,	7:	6,	10
2,	5:	5,	7	4,	5:	6,	9
2,	10:	5,	5	4,	6:	5,	10
2,	10:	7,	7	4,	9:	5,	8
2,	7:	10,	10	4,	5:	4,	5
2,	2:	6,	6	4,	4:	10,	10
2,	4:	4,	6	4,	8:	9,	9
2,	7:	3,	6	5,	5:	6,	8
2,	8:	4,	4	5,	8:	10,	10
2,	7:	2,	10	6,	8:	8,	10
2,	3:	7,	8	6,	9:	7,	7
3,	9:	5,	7	9,	10:	9,	10
3,	8:	4,	5	-	-	-	-

Note: The order in each pair is immaterial, and the order of the pairs themselves is also immaterial (so 9,8:2,4; 8,9:2,4; 9,8:4,2; 8,9:4,2; 4,2:8,9; 4,2:9,8, and 2,4:9,8 are all equivalent to the first entry in the table, 2,4:8,9)

Table 4. Two-tailed 95% confidence limit; these are the group sizes that are just significant at the exact 95% confidence limit, but in the normal approximations are just not significant (*z* < ±1.65 to 3 sf., but for all *z* > ±1.520)

2,	2:	2,	5	2,	6:	3,	6	3,	3:	6,	6	4,	6:	4,	10
2,	2:	2,	9	2,	6:	3,	9	3,	3:	9,	9	4,	7:	6,	9
2,	2:	3,	3	2,	6:	5,	5	3,	4:	9,	10	4,	7:	7,	10
2,	2:	3,	5	2,	6:	5,	8	3,	5:	3,	6	4,	8:	9,	10
2,	2:	4,	6	2,	6:	6,	6	3,	5:	7,	10	4,	9:	8,	8
2,	2:	5,	8	2,	6:	6,	9	3,	6:	3,	9	4,	10:	4,	10
2,	3:	2,	10	2,	6:	8,	8	3,	6:	5,	5	4,	10:	6,	8
2,	3:	3,	9	2,	6:	10,	10	3,	6:	6,	10	5,	5:	6,	9
2,	3:	4,	4	2,	7:	3,	3	3,	6:	7,	7	5,	6:	6,	8
2,	3:	5,	9	2,	7:	4,	4	3,	6:	7,	8	5,	8:	7,	8
2,	3:	6,	8	2,	7:	6,	8	3,	7:	7,	7	5,	9:	6,	7
2,	3:	8,	9	2,	8:	2,	8	3,	7:	8,	9	5,	9:	8,	9
2,	4:	2,	8	2,	8:	6,	10	3,	8:	3,	8	5,	9:	9,	9
2,	4:	5,	6	2,	8:	7,	7	3,	8:	6,	6	5,	10:	8,	8
2,	4:	5,	10	2,	9:	4,	5	3,	8:	7,	7	6,	7:	6,	9
2,	4:	6,	10	2,	9:	5,	6	3,	8:	10,	10	6,	7:	9,	9
2,	5:	3,	10	2,	9:	6,	9	3,	9:	5,	6	6,	8:	6,	10
2,	5:	4,	7	2,	9:	8,	10	3,	9:	5,	8	6,	10:	10,	10
2,	5:	4,	8	2,	10:	3,	3	3,	9:	9,	9	7,	7:	10,	10
2,	5:	5,	7	2,	10:	3,	4	3,	10:	4,	7	7,	8:	7,	9
2,	5:	6,	7	2,	10:	4,	10	4,	4:	6,	8	7,	10:	8,	9
2,	5:	8,	10	2,	10:	5,	5	4,	5:	5,	8	8,	10:	9,	10
2,	5:	2,	8	3,	3:	5,	8	4,	5:	7,	10	-	-	-	-

Note: As described in Table 3, the order in each pair is immaterial, as is the order of the pairs themselves

Conclusion

The explanatory factorial approach taken here can be directly extended to any of the large family of tests in which the full extent of all possible test statistics is created and the most extreme tail of that distribution defined as the critical region. This includes all the WMWt family of tests (e.g. Jonckheere-Terpstra Test, Kruskal-Wallis Test), Kolmogorov-Smirnov-type tests (e.g. Conover test, Birnbaum-Hall Test), and Pitman permutation tests (Conover, 1999). The generic approach is:

1. Separate the experimental subjects across all factors and establish the test statistic for the two factors of experimental interest
2. Sum the test statistics across all the nuisance factors to get a combined test statistic

3. Derive the probability mass function for the combined statistic under the null hypothesis
4. Establish if the combined test statistic from step 2 is in the extreme tail of the distribution derived in step 3

This method also has direct extension to factors that hold 3 or more states – if, for example, one ran an experiment with three or more different strains of animal, in several age groups of clinical patients, or with clinical results from three or more different hospitals, it would then be possible to form a combined *U*-value by summing the three separate results. This could be compared with a similarly-derived probability mass function for this value under the null hypothesis.

There are weaknesses in the method. If one sex responds in a quantitatively different manner to the other (interaction of treatment with another factor), this is not measured. Subsetting the data on the nuisance factor and doing separate analyses on U_1 and U_2 might make one suspect interaction. Ties are not obviously incorporated in this simple formulation, although the method for correcting for ties in the derivation of U_1 and U_2 should be applicable to give a U_C that is unaffected by ties in U_1 and U_2. We are currently developing this approach to ordinal contingency table data, in which very extensive ties are usual.

There are other approaches that might be adopted to achieve the same ends. Substituting normal scores for the ranks and then using a procedure such as ANOVA would be one. A Shirley test approach (Williams, 1986) of using the ranks themselves as though they were interval data and using a parametric procedure might be possible. When the test statistic for each subset of the data approximates to an established distribution (at it does in the Kruskal-Wallis test with the X^2 distribution), then combining those individual statistics may be possible for a test of the complete data set. Fisher's combined probability test and Weighted Z methods are inappropriate as they assume the probability for each table is uniformly distributed on the interval [0, 1], but their *p* values can only hold discrete values. However, the approach we advocate has these advantages:

1. It is valid for very small groups; with 2 sexes, a 2 factor study with just two subjects for each sex and treatment can give significant results
2. It is robust with any distribution in the data; the groups do not even have to be drawn from the same distribution family, simply independently of each other

3. Experimenters can clearly comprehend the rationale behind the test
4. The meaning of the result is clear
5. It is computationally simple

There are numerous practical situations in which experimenters may want to use a non-parametric test such as the WMWt, but there exists in their study design an unavoidable ‘nuisance factor’ which precludes the simple application of the test. Animal experiments, human clinical trials data, and psychological tests that include data from both sexes or can be age-stratified are all examples of large classes of such experiments. It is very rare that one can unequivocally exclude sex or age as a possible latent factor in such experiments, so it would be prudent to adopt as routine such methods as these. In the field in which one of us works – toxicology – it is common for experiments to be conducted with small groups sizes (*n* = 3 or 4) for studies involving primates or dogs (there are ethical objections to primate experiments with large group sizes). Doing the analyses on the sexes separately markedly limits the power of such experiments. However, doing tests including sex as a factor in the analysis of the complete data sets has a substantial beneficial effect on the power of such experiments. Simulation shows that with group sizes of 3 and using data from both sexes and all dose groups the power then approaches that of larger studies with group sizes of 10 in which the data for the sexes is not combined (Holland, 2011).

References

Berkley K. J., Zalcman, S. S., & Simon, V. R. (2006). Sex and gender differences in pain and inflammation: A rapidly maturing field. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology, 291*(2), R241-R244. doi: 10.1152/ajpregu.00287.2006

Blair, R. C., Higgins, J. J., & Smitely, W. D. G. (1980). On the relative power of the *U* and *t* tests. *British Journal Mathematical and Statistical. Psychology, 33*(1), 114-120. doi: 10.1111/j.2044-8317.1980.tb00783.x

Campbell, R. C. (1974). *Statistics for biologists* (2nd ed.). London: Cambridge University Press.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York, NY: John Wiley and Sons.

Di Bucchianico, A. (1999). Combinatorics, computer algebra and the Wilcoxon-Mann-Witney test. *Journal of Statistical Planning and Inference, 79*(2), 349-364. doi: 10.1016/S0378-3758(98)00261-4

Holland, T. (2010). *A study of the methods used in toxicological pathology* (Fellowship thesis). Retrieved from Royal College of Veterinary Surgeons' Library, London (T/712).

Sprent, P., & Smeeton, N. C. (2001). *Applied non-parametric statistical methods* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Williams, D. A. (1986). A note on Shirley's nonparametric test for comparing several dose levels to a non-zero control. *Biometrics, 42*(1), 183-186. doi: 10.2307/2531254

Appendix 1

Matlab Program Calculating f₁(*m*, *n*, *r*) for specified *m*, *n*

$f_1(m, n \mid r)$ is specified for given inputs m, n by fvector[i] = $f_1(m, n \mid i - 1)$.

```
function [ fvector ] = U1( ntreat, ncntrl )

m=ntreat;
n=ncntrl;

farray=zeros(m+1,n+1,m*n+1);

farray(1,1,1)=1;

fvector=farray(1,1,1);

for j=2:(m+1)
    farray(j,1,1)=1;
end;

for k=2:(n+1)
    farray(1,k,1)=1;
end;

for k=2:(n+1)
    for j=2:(m+1)
        for l=(1:(k-1)*(j-1)+1)
            sum=0;
            for h=1:min(l,k)
                sum=sum+farray(j-1,h,l-h+1);
            end;
            farray(j,k,l)=sum;
        end;
    end;
end;

fvector=zeros(m*n+1,1);

for l=1:(m*n)+1
    fvector(l,1)=farray(m+1,n+1,l);
end;

end
```

Appendix 2

Matlab Program Calculating $\mathrm{f}_2(m, n; p, q \mid k)$

$\mathrm{f}_2(m, n; p, q \mid r)$ is specified for given inputs m, n, p, q by fvector[i] = $\mathrm{f}_2(m, n; p, q \mid i - 1)$.

```
function [ fvector ] = U2( ntreat1, ncntrl1, ntreat2, ncntrl2 )

m1=ntreat1;
n1=ncntrl1;
m2=ntreat2;
n2=ncntrl2;

m=max(m1,m2);
n=max(n1,n2);

farray=zeros(m+1,n+1,m*n+1);

farray(1,1,1)=1;

fvector=farray(1,1,1);

for j=2:(m+1)
farray(j,1,1)=1;
end;

for k=2:(n+1)
    farray(1,k,1)=1;
end;

for k=2:(n+1)
    for j=2:(m+1)
        for l=(1:(k-1)*(j-1)+1)
            sum=0;
            for h=1:min(l,k)
                sum=sum+farray(j-1,h,l-h+1);
            end;
            farray(j,k,l)=sum;
        end;
    end;
end;

fvector1=zeros(m1*n1+1,1);
```

```
for l=1:(m1*n1)+1
    fvector1(l,1)=farray(m1+1,n1+1,l);
end;

fvector2=zeros(m2*n2+1,1);

for l=1:(m2*n2)+1
    fvector2(l,1)=farray(m2+1,n2+1,l);
end;

fmatrix=zeros(m1*n1+1,m2*n2+1);

for j=1:m1*n1+1
    for k=1:m2*n2+1
        fmatrix(j,k)=fvector1(j,1)*fvector2(k,1);
    end;
end;

fvector=zeros(m1*n1+m2*n2+1,1);

for j=1:m1*n1+m2*n2+1
    sum=0;
    for k=max(1,j-m1*n1):m2*n2+1
        if k-j>0
            break;
        end;
        sum=sum+fmatrix(j-k+1,k);
    end;
    fvector(j,1)=sum;
end;

end
```