Exact and Pseudo P-values in the Wilcoxon Unpaired Test with Ties

Takeshi Otsu

Abstract

This paper investigates the discrepancy between the exact and the pseudo p-values of the Wilcoxon unpaired test statistics for several tie patterns. The findings are summarized as follows. First, the discrepancy between the exact and the pseudo p–values tends to be less than 5% when the sample size is greater than 15 for each sample set and the size of the ties takes 15% at most in the sample. Secondly, the exact and the pseudo p-values are very similar at the significance levels from 0.01 to 0.1. Since these levels are frequently used in practice, the statistical testing is less likely to be misleading. Thirdly, existence of mid ranks or a large size of the ties beyond 20% in the sample deteriorates approximations by the pseudo p-values. Finally, the exact p-values are easily obtained with a simple algorithm on a worksheet application when each sample size of two data groups is less than or equal to 16.

1. Introduction

The Wilcoxon unpaired test is widely used in the biomedical sciences and the social sciences, where we need to perform a statistical test whether or not two sets of samples come from the same distribution. Although it was named after Wilcoxon (1945), many other scientists proposed similar tests. Kruskal and Wallis (1952) and Kruskal (1957) gave a useful historical review, and we can find an extensive bibliography in Jacobson (1963).

The basic idea is simple and explained in a standard text book, such as Lehmann (1998). To compute the Wilcoxon's rank-sum statistic, we simply rank all the combined sample points, and sum up ranks for one of the sample sets. To compute the significance probability (or p-value) under the null distribution, we just count the number of combinations of the observed points that give rise to values less than or equal to the Wilcoxon statistic, and divide it by the number of all the possible combinations of the combined sample points. Although the idea is simple, it is time-consuming and sometimes infeasible to calculate the numerator of the significance probability. This computational burden spawns a lot of studies for approximation and efficient computational methods. Tabulation is one way to deal with the computational burden. Jacobson (1963) and Verdooren (1963) used the method proposed by Fix and Hodges (1955) to tabulate critical values, which they compared with those of the earlier studies. Verdooren (1963) presented tables of critical values of the Wilcoxon's rank-sum statistic at significance levels of 0.001, 0.005, 0.01, 0.025, 0.05 and 0.10 for sample sizes of 25 or less. Jacobson (1963) gave the critical-value tables for significance levels of 0.005, 0.01, 0.025, and 0.05 when sample sizes were 29 or less. In practice, however, the computed statistics at hand may not be covered in the tables. Therefore, it would be convenient to compute p-values for statistical testing.

Another strand in the literature explores the accuracy of approximation. Mann and Whitney (1947) rigorously examined statistical properties of the Wilcoxon statistic, or its equivalent called the Mann-Whitney U statistic, to prove that its limit distribution is normal when there is no tie in each sample. Kruskal and Wallis (1952, pp. 591-595), Hodges, Ramsey, and Wechsler (1990) and Lehmann (1998, p. 17) reported the accuracy of the normal approximation with a continuity correction. Generally speaking, the approximation is not good for small p–values. Kruskal and Wallis (1952) found an anomaly that the continuity correction made the approximation worse when the p-value was less than or equal to 0.02. Hodges, Ramsey, and Wechsler (1990, p. 251) pointed out that their approximation did not correct the kurtosis error that tended to increase the p-value, and the continuity correction also increased the p-value. Therefore, without the kurtosis correction, the continuity correction came to substantially overestimate the true value.

Taking advantage of the symmetry of the U statistic, Fix and Hodges (1955, p. 311) obtained the Edgeworth approximation formula with the first six moments and a continuity correction, which included the kurtosis-error correction. Despite its accuracy, the complexity of the formula limits its usefulness. Hodges, Ramsey, and Wechsler (1990) proposed a simplified formula, and showed that both of the formulae gave a good approximation to the exact p-values, as long as each sample size was 12 or more. Jacobson (1963) compared the normal approximation to critical values with the exact values, showing its poor performance especially at lower significance levels. Verdooren (1963) examined the accuracy of the normal and the Edgeworth approximations when one of the sample sets had observations of 25 or more, and showed the Edgeworth expansion gave a better accuracy.

By and large, the literature indicates that the exact significance probability should be used when sample sizes are less than or equal to 25, and the Edgeworth expansion with kurtosis-error and continuity corrections works reasonably well when sample sizes are more than 25. Fortunately, technological advance allows us to compute the exact p-values. Particularly, the computational method proposed by Fix and Hodges (1955) is easily programmed and efficiently works even with worksheet softwares. One of the remaining problems is how to deal with samples including ties. Since all the argument up to here only applies to untied samples, we need further investigation on effects of tied observations.

When ties are present in the samples, the mean of the Wilcoxon's rank-sum statistic remains same as that without ties, but the variance and higher moments need modification. Kruskal (1952) and Kruskal and Wallis (1952) derived the mean and variance. Lehman (1961) investigated accuracy of the normal approximation with corrections for continuity and ties. It examined six artificial data sets, one of which had no ties and others had different patterns of ties, and found very poor approximations when samples were heavily tied. Further, tail probabilities were badly approximated because the tails of the discrete distribution were thinly populated and the sample sizes were small. Continuity corrections were advisable for better approximations. Interestingly, the approximation at the significance level of 0.01 was better than that at the level of 0.05 or 0.1. It also confirmed the working guide, suggested by Kruskal and Wallis (1952, p. 587), that the correction for ties in the normal approximation would not change estimated values of significance probability by more than 10%, with sample sets of ten or fewer observations if they involved ties not more than one-fourth of the sample points. Klotz (1966), however, concluded that the erratic nature of the statistic values ruled out any smooth approximation. It derived conditional moments through the fourth order under the null distribution for the Edgeworth expansion with tie corrections. It compared the normal and the Edgeworth approximations with the exact p-values, and found the irregularity of the distribution when the sample size was less than or equal to 12 for each group and 22 in all at most. Therefore, it would be necessary to compute the exact probability values.

The purpose of this paper is to compare exact significance probabilities with ties to the pseudo probabilities computed assuming no ties. In general, observations with ties make computational burden of p-values heavier than those without ties. Therefore, if researchers need to work in a limited computational resource environment such that only worksheet softwares are available, it may be prohibitive to obtain the exact p-values for tied samples. Even statistical software packages may not produce exact p-values with ties. Bergmann, Ludbrook, and Spooren (2000) reviewed 11 commercial statistical packages to find that 9 packages took ties into consideration in large-sample approximations, but only four packages gave exact p-values for tied samples. Worse, documentations of these packages tended to inadequately describe algorithms and correction methods. Thus, statistical testing based on these p-values may lead up to a wrong conclusion in small samples. If exact p-values are not available, it might be better to use pseudo p-values under the false assumption of no ties than the large sample approximations. Verdooren (1963, p. 179) conjectured that we might still use the critical values without ties if the size of the ties were not very large. Fortunately, the method proposed by Fix and Hodges (1955) works efficiently even with worksheet softwares to compute the exact p-values with untied observations. Then, researchers need to understand to what extent a pseudo p-value deviates from the corresponding exact p-value, so that they avoid reaching wrong statistical conclusions.

To investigate the discrepancy between the exact and the pseudo p-values, we conduct simulations with artificial data with several patterns of ties considered in the literature. The findings are summarized as follows. First, the discrepancy between the exact and the pseudo p-values tends to be less than 5% when the sample size is greater than 15 for each sample set and the size of the ties occupies 15% at most in the combined sample. Secondly, the exact and the pseudo p-values are very similar at the significance levels from 0.01 to 0.1. Since these levels are frequently used in practice, the statistical judgment is less likely to be mislead. Thirdly, existence of mid ranks or a large size of the ties beyond 20% in the sample deteriorates approximations by the pseudo p-values. Finally, we find it feasible to compute the exact p-values with a worksheet application such as Excel (Microsoft) when each sample size is less than or equal to 16. In the following section, we briefly explain how to compute the exact and the pseudo p-values. In Section 3, we present simulation results to see how the pseudo p-values deviate from the exact values. The final section is allocated to discussion of the future research topics.

2. Exact and pseudo p-values

Suppose we have two independent random samples $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ from two populations with unknown cumulative distribution functions, F and G, respectively. Then, we have interest in testing the hypothesis $H_0 : F(t) = G(t)$ against the one-sided alternative hypothesis $H_A : F(t) \geqq G(t)$ and $F(t) \neq G(t)$ for some t. Wilcoxon (1945) proposed the statistic

$$W = \sum_{i=1}^{m} r_i \tag{1}$$

where r_i , · ,r_m are the ranks of X_1 ,· · ·, X_m in the combined sample. Equivalently, Mann and Whitney (1947) used the following statistic to show its asymptotic normality as m and n go to infinity:

$$U = \sum_{i=1}^{m} r_i - \frac{m(m+1)}{2} \tag{2}$$

Suppose the sample set of X_1 ,· · ·, X_m gives a smaller mean of the rank sum than the other sample set. Then, the lower tail probability under the null distribution is written

as Prob ($U \leqq t | H_o$), which we call the p-value. It is imperative to compute the accurate p-value for statistical testing. As shown in Lehmann (1998, pp. 12-13), under the null hypothesis, these statistics have symmetric distributions about m(m + n+1)/2 when there is no tie in the sample. Then, once we know the exact one-sided tail probability, we can use it for one-sided tests as well as two-sided tests. Although the presence of ties makes the null distribution asymmetric and dependent on the tie patterns, Klotz (1966) showed the statistics would have a symmetric null distribution conditional on a tie pattern if two samples have the same size (m = n) or if the tie patterns are symmetric. Thus, it might be still useful to use the doubled p-values for two-sided tests in practice, at least approximately.

It is costly to tabulate the p-values as a function of *m, n,* and *t,* or critical values at each significance level without ties, and almost impossible with ties. In the literature through the middle of 1980's, several algorithms were proposed to compute the p-values under the constraints of storage sizes and processing speeds of computers. For example, Klotz (1966) extended a recurrence relation of probabilities of the statistic U of

Mann and Whitney (1947) into an algorithm to enumerate the exact null distribution with ties. Hill and Peto (1971) presented an algorithm that recursively calls a subroutine to make storage constraints unbinding. Mehta, Patel, and Tsiatis (1984) applied a network algorithm to reduce running times. Soms (1977) also proposed a similar algorithm. All these algorithms require a lot of do-loop processing, and would be suitable for programming in Fortran or C language. Since worksheet applications are not good at do-looping, however, it may not be practical to compute the exact p-values when only a worksheet application is available.

To compute exact p-values when ties are present, we use the algorithm presented by Richards and Byrd (1996), which is logically simple and materialized in Fortran language, because it requires less do-loop processing than other preceding algorithms and comfortably works on worksheet applications. The basic idea of Richards and Byrd (1996) is as follows. First, we pick up one of the two sample sets, which has a smaller mean of the ranks. Then, we trade equal numbers of observations between the samples to identify combinations to produce rank sums lower than or equal to the rank sum of the chosen sample group. If we have two sample sets, each of which has10sample points, the possible number of tradings amounts to 1023 ($= \sum_{i=1}^{10} {}_{10}C_i$). Thus, the algorithm computes 2024 summations for the two sample sets, since no summations are necessary in cases of one-for-one (${}_{10}C_1 = 10$) and ten-for-ten (${}_{10}C_{10} = 1$) tradings: $2024 = (1023-11) \times 2$. Since all the possible combinations of the rank sums are 184, 756 (${}_{20}C_{10}$), it substantially reduces computational burdens. To carry out investigations here, we modified the Fortran program so that it could handle two data sets of 42 samples in all, 21 samples for each at most, with double precision. Further, in the experimental process, we translated the original Fortran program into the Visual Basic program in Excel (Microsoft) to see to what extent it would be feasible to obtain exact p-values only with the worksheet application. We checked the program against data in the literature such as Bergmann, Ludbrook, and Spooren (2000, p. 74) and Lehman (1961, p. 294).

Turning to the pseudo p-values, that is, the p-values computed without taking account of ties, we use the method proposed by Fix and Hodges (1955). Let m . n positive integers, and A(u,m,n) the number of combinations in which it is possible to choose exactly m nonnegative integral summands, none greater than n whose sum does not exceed *u*. If π (*u*, *m*, *n*) denotes the distribution function of U in eq. (2), we can write

$$Prob(U \leqq u \mid H_0) = \pi(u,m,n) = \frac{A(u,m,n)}{\binom{m+n}{m}} \tag{3}$$

To compute $A(u, m, n)$, we consider the number of combinations without the constraints of none greater than n. That is,

$$A_0(u,m) = A(u,m,\infty) \tag{4}$$

where $A_0(u, m)$ and $A(u, m, n)$ are 0 when u < 0. When all the variables of summation are integers, we have

$$A(u,m,n) = A_0(u,m) - \sum_{t=n+1}^{\infty} A(u-t,m-1,t) \tag{5}$$

Using this relation, Fix and Hodges (1955) expressed the restricted partition function A in terms of the unrestricted partition functions as follows:

$$A(u,m,n) = \sum_{k=0}^{\infty}(-1)^k A_k\left(u - kn - \frac{1}{2}k(k+1), m-k\right) \tag{6}$$

where

$$A_k(u, m-k) = \sum_{v=0}^{u} a_0(v, k) A_0(u-v, m-k) \qquad (7)$$

and

$$a_o(u, m) = A_0(u, m) - A_0(u-1, m) \qquad (8)$$

The boundary conditions are $A_0(0, m) = 1$ and $A_0(u, 1) = u + 1$. As noted by Fix and Hodges (1955), Leonhard Euler, a Swiss mathematician, studied in the early 1900's the function a_0 that gives the number of combinations partitioning the exact value u into m parts. We programmed an algorithm to compute a_0 in Visual Basic included in Excel, and tabulated $A_0(u, m)$ on the worksheet for $u \leqq 100$ and $m \leqq 12$, using the relation of eq. (8). We checked the values with those of Fix and Hodges (1955) to confirm our algorithm worked properly. To reduce computational costs, we use the following convenient recursive relation (Fix and Hodges, 1955, p. 303) for $101 \leqq u \leqq 1000$ and $13 \leqq m \leqq 30$:

$$A_0(u, m) = A_0(u, m-1) + A_0(u-m, m) \qquad (9)$$

This recursion formula can be verified using 0-1 sequences examined in Mann and Whitney (1947). In the Appendix A, we give a simple example to confirm the validity of this relation. We used eq. (6) to compute $A(u, m, n)$ and then calculated the pseudo p-value with eq. (3). It took less than a few seconds to obtain the pseudo p-value with the worksheet software.

3. Simulation Results

Before we examine simulation results, we briefly report computational time to obtain exact p-values with Excel 2000 (Microsoft), using the algorithm of Richards and Byrd (1996). We used a laptop computer with Pentium III processor with 500 MHz (megahertz) and 256 MBs (megabytes) memories. When each sample had 16 data points, it took 10 seconds. The computation was getting slower and slower with a sample size of more than 16. When each data set had 17 samples, it took one minute and a half. The computational time tripled to 5 minutes when the sample size was 18, and further tripled to 16 minutes when it was 19. Then, it jumped up to 70 minutes for the sample size of 20. Therefore, it would be practical to compute exact p-values with the worksheet software when the sample size is less than or equal to 16 for each sample.

In our simulation, we consider tie patterns in Klotz (1966, pp. 777-779). If we write a tie pattern as (1, 2, 1), it means 1, 2, 2, 4 in rank. It is converted to 1, 2.5, 2.5, 4 in mid rank, the averaged rank of the tied samples, to compute the Wilcoxon statistic. When we write (1, 2, 1) × 2, it indicates (1, 2, 1, 1, 2, 1), that is, 1, 2.5, 2.5, 4, 5, 6.5, 6.5, 8 in mid rank. According to the literature that studied the precision of the normal and the Edgeworth approximation, the approximation to the tail probabilities was worse than that to the middle-part probabilities (see, for example, Kruskal and Wallis 1952, Lehman 1961, Jacobson 1963, Hodges, Ramsey, and Wechsler 1990, and Lehmann 1998). Thus, we focus on the tailed probabilities from 0.01 to 0.1, which correspond to the significance levels frequently used in practice.

Table1shows simulation results of sample sizes less than or equal to 16. When the statistic takes the mid-ranked value such as 17.5, 21.5, 38.5 and so forth, the difference rate, that is, the rate of change from the exact p-value to the pseudo p-value, tends to be more than 19%. The largest discrepancy 69% is observed for the statistic 36.5 in CASE 4 though its p-value is so small that it may not be used as a significance level in practice. On the contrary, the difference rate is less than 10% at the significance levels more than 0.01 when the statistic is not mid-ranked. This is partly because the mid-ranked statistic is truncated to a nearest smaller integer to compute the pseudo p-value. Then, the pseudo p-value underestimates the exact value. Although it is somewhat subjective to determine what error rates should cause a serious problem, Hodges, Ramsey, and Wechsler (1990, p. 251) claimed that a 15% would be too large and not trivial. In their view, even if the sample sizes are small, it can be said that the pseudo values give a good approximation except when the statistic is mid-ranked. The largest difference in absolute value is 0.024 of the statistic 21.5 in CASE1,which is about 18% of the exact value.

Turning to Table 2, the difference rates are still bumpy across the statistics, but mostly smaller in absolute value than those in Table 1. This suggests that a larger sample size ameliorates the pseudo-value approximations. Interestingly, the approximation is better for the mid-ranked values than the just-ranked ones in some cases, such as in the statistic of 106.5 in CASE 7 or 78.5 in CASE 8. The difference rates are less than 15% for the p-values ranging from 0.01 to 0.1. The largest difference is 0.011 when the statistic is 95.0 in CASE 8,8% of the exact value that is less than half of the value, 18% in CASE 1 mentioned above. Therefore, the larger sample size in all, the better the approximation. In terms of tie patterns, when the size of the ties has a small share in the whole sample, the accuracy of approximation is substantially improved. In CASE 10, the tie takes about 8% in the combined sample, while it occupies20%or more in other cases of Table 2. This is consistent with the conjecture of Verdooren (1963, p. 179) mentioned before.

In Table 3, we find that the mid ranks in higher ranks cause the difference rates more bumpy than those in lower ranks. The data of CASE 15 are heavily tied in higher ranks: there are six 16th's and seven 22nd's in rank. In contrast, those of CASE 16 are heavily tied in lower ranks. Comparing these cases, the difference rates go up and down more sharply in CASE 15 than in CASE 16. The tie patterns of CASE 15 and CASE 16 are similar to CASE 8 and CASE 9 in Table 2: the size of the ties monotonically increases or decreases. In these cases, the approximation is better for the mid-ranked values than the just-ranked ones. The comparison of CASE 12 with CASE 13 indicates that the mid ranks may substantially deteriorate the approximation by the pseudo p-values, annihilating the benefit of a smaller size of the ties, only 7% in the whole sample. Generally, larger the sample sizes, lower the difference rates. However, the pseudo p-value seems give a better approximation when both data sets increase their sample sizes in a balanced manner. The observations are evenly tied in CASE 5 of Table 1, CASE 10 of Table 2, CASE 12 and CASE 13 of Table 3. When the number of observations is same for both sample sets, the difference rates decrease as more observations are added. When the number of samples increases in one of the sample sets less than in the other as in CASE 12 and CASE 13, however, the approximation does not improve or even deteriorate. The largest difference in absolute value, 0.008, is observed for the statistic 161.5 in CASE 12, at approximately 10% significance level.

Finally, in Table 4, the difference rates are less than 5% in many cases when both sample sizes are more than 15 except CASE 19. Although the error rates are relatively large in CASE 19, they are still less than 10% in spite of the mid-ranked statistics, comparable with CASE 10 and better than CASE 5 and CASE 12 that have similar tie patterns. Thus, if each of the sample sets has more than 15 observations, the pseudo p-values will give close approximations to the exact p-values. The largest difference is 0.005 in absolute value, when the statistic is 287.5 in CASE 18. It is only 5% of the exact value. It is noted that the size of the ties in Table 4 occupies 15% at most in the combined sample.

4. Discussion

This paper investigates the discrepancy between the exact and the pseudo p-values of the Wilcoxon unpaired test statistics for several tie patterns. The main findings are summarized as follows. First, we find that the discrepancy between the exact and the pseudo p-values tends to be less than 5% when the sample size is greater than 15 for each sample set and the maximum size of the ties takes 15% in the sample. Secondly, the exact and the pseudo p-values are very similar at the significance levels from 0.01 to 0.1. Since these levels are frequently used in practice, the statistical testing is less likely misleading. Thirdly, the pseudo p-values may not approximate the exact values very well when the mid ranks exist or when the size of the ties occupies greater than or equal to 20% in the whole sample. Finally, it is feasible to obtain the exact p-values with a worksheet application when each sample size is less than or equal to 16.

Several caveats are in order. We focus on the p-values from 0.1 to 0.01, which correspond to the significance levels frequently used in practice, because the literature expects the middle-part probabilities are approximated fairly well. However, we might need to investigate the extent of approximation in the middle part for completion. Further, we do not examine degrees of asymmetry when ties exist. Thus, in some cases, it may not be appropriate to use doubled p-values for two-sided tests. Finally, it would be useful to develop an algorithm to compute exact p-values with tied samples, which requires less do-loop processing. Since worksheet softwares are prevailing, a do-loop-free algorithm would be convenient to conduct the Wilcoxon unpaired test even in a poor computing environment. These are left for the future research topics.

Acknowledgement: It is gratefully acknowledged that this work is supported by JSPS (Japan Society for the Promotion of Science) KAKENHI (17530223).

Appendix A : Recursion Fromula

In this appendix, we give a simple example to confirm the relation of eq. (9). First, we recursively apply eq. (8) to itself to obtain the following equation:

$$A_0(u,m) = \sum_{j=0}^{m-1} a_0(u-j,m) + A_0(u-1,m) \tag{10}$$

Suppose we have $u = 3$ and $m = 2$. Then, we have

$$A_0(3,2) = a_0(3,2) + a_0(2,2) + A_0(1,2) \tag{11}$$

Now we note, letting u_i the ith element of U in eq. (2),

$$u_i = r_i - i \tag{12}$$

where $r_1, \cdot, r_m$ rm are the ranks of $X_1, \cdots, X_m$ in the combined sample as in eq. (1). Then, $u_1 = 3$ and $u_2 = 2$ in the example here. Since $u_1 = 1+2 = 0+3, a_0(3,2)$ is equal to 2. In terms of the rank r_i, the summations are 2 + 4 and 1 + 5. That is, the sample points of the sample set X take the ranks of 2 and 4, or 1 and 5. Let us denote 0 to indicate the location of the samples from the X, andlfor the samples from the other set. Then, we can construct the 0-1 sequences as follows:

$a_0(3,2): u_1 = 1+2,\ W = 2+4$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	1	0	1	0	1	1	$\cdots$

In this table, the first ’1’ precedes ’0’ twice, or two 0’s, and the second one precedes ’0’ once. Therefore, the ’1’ precedes ’0’ three times in all, which corresponds to the number of u_1 (= 3). Similarly, we have

$a_0(3,2): u_1 = 0+3,\ W = 1+5$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	0	1	1	1	0	1	$\cdots$

$a_0(2,2): u_2 = 1+1,\ W = 2+3$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	1	0	0	1	1	1	$\cdots$

$a_0(2,2): u_2 = 0+2,\ W = 1+4$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	0	1	1	0	1	1	$\cdots$

Suppose we cut the part of the 0-1 sequence left to the first ’0’ and slide the remaining sequence to the left. Then we have the following 0-1 sequences:

$a_0(1,1): u_3 = 1,\ W = 2$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	1	0	1	1	1	1	$\cdots$

$a_0(3,1): u_1 = 3,\ W = 4$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	1	1	1	0	1	1	$\cdots$

$a_0(0,1) : u_1 = 0,\ W = 1$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	0	1	1	1	1	1	$\cdots$

$a_0(2,1) : u_2 = 2,\ W = 3$

rank (r_i)	1	2	3	4	5	6	$\cdots$
location of X_i	1	1	0	1	1	1	$\cdots$

Consequently, we have converted m to $m-1$. Note we renumbered the subscripts of u_i's in descending order of their values. We may rewrite eq. (11) as:

$$\begin{aligned} A_0(3,2) = {} & a_0(3,1) + a_0(2,1) + a_0(1,1) + \\ & a_0(0,1) + A_0(1,2) \end{aligned} \tag{13}$$

or, compactly,

$$A_0(3,2) = A_0(3,1) + A_0(1,2) \tag{14}$$

In general, we have the recurrence relation of eq. (9) in Section 2.

References

Bergmann, R., J. Ludbrook, and W.P.J.M. Spooren, 2000, "Different Outcomes of the Wilcoxon-Mann-Whitney Test From Different Statistics Packages," The American Statistician, 54 (1), 72-77.

Dinneen, L.C., and B.C. Blakesley, 1973, "Alogrithm AS 62: A Generator for the Sampling Distribution of the Mann-Whitney U Statistic," Applied Statistics, 22 (2), 269-273.

Fix, E., and J.L. Hodges, Jr, 1990, "Significance Probabilities of the Wilcoxon Test," Annals of Mathematical Statistics, 26, 301-312.

Hill, I.D. and R. Peto, 1971, "Algorithm AS 35: Probabilities Derived from Finite Populations," Applied Statistics, 20 (1), 99-105.

Hodges, J.L., Jr., P.H. Ramsey, and S. Wechsler, 1990, "Improved Significance

Probabilities of the Wilcoxon Test," Journal of Educational Statistics, 15 (3), 249-265.

Jacobson, J.E., 1963, "The Wilconxon Two-Sample Statistic: Tables and Bibliography," Journal of the American Statisitcal Association, 58 (304), 1086-1103.

Klotz, J.H., 1966, "The Wilcoxon, Ties, and the Computer," Journal of the American Statisitcal Association, 61 (315), 772-787.

Kruskal, W.H., 1952, "A Nonparametric Test for the Several Sample Problem," Annals of Mathematical Statistics, 23 (4), 525-540.

Kruskal, W.H., 1957, "Historical Notes on the Wilcoxon Unpaired Two-Sample Test," Journal of the American Statisitcal Association, 52, 356-360.

Kruskal, W.H., and W.A. Wallis, 1952, "Use of Ranks in One-Criterion Variance Analysis," Journal of the American Statisitcal Association, 47 (260), 583-621.

Lehman, S.Y., 1961, "Exact and Approximate Distributions for the Wilcoxon Statistic with Ties," Journal of the American Statisitcal Association, 56 (294), June, 293-298.

Lehmann, E.L., 1998, Nonparametrics: Statistical Methods Based on Ranks, Springer.

Mann, H.B., and D.R. Whitney, 1947, "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other," Annals of Mathematical Statistics, 18 (1), 50-60.

Mehta, C.R, N.R. Patel, and A.A. Tsiatis, 1984, "Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data," Biometrics 40, 819-825.

Richards, L.E., and J. Byrd, 1996, "Algorithm AS 304: Fisher's Randomization Test for Two Small Independent Samples," Applied Statistics, 45 (3), 394-398.

Soms, A.P., 1977, "An Algorithm for the Discreate Fisher's Permutation Test," Journal fo the American Statistical Association, 72 (359), September, 662-664.

Verdooren, L.R., 1963, "Extended Table of Critical Values for Wilcoxon's Test Statistic," Biometrika, 50 (1 and 2), 177-186.

Wilcoxon, F., 1945, "Individual Comparisons by Ranking Methods," Biometrics Bulletin, 1 (6), 80-83.

Table 1 Discrepancy between exact and pseudo p - values: $m + n \leqq 16$

CASE 1 tie pattern: (1,1,2,1,1,2,1,1), sample sizes: (m, n)=(5, 5)

Wilcoxon	exact	pseudo	difference rate (%)	difference
statistic	p - value	p - value	(log_e(exact/pseudo)×100)	(pseudo - exact)
15.0	0.003968	0.003968	0.0064	0.000000
16.0	0.007937	0.007937	-0.0062	0.000000
17.5	0.023810	0.015873	-40.5485	-0.007937
19.0	0.047619	0.047619	0.0001	0.000000
20.0	0.071429	0.075397	5.4061	0.003968
21.5	0.134921	0.111111	-19.4159	-0.023810

CASE 2 tie pattern: (1,1,1,2,1,1,2,1), sample sizes: (m, n)=(5, 5)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
15.0	0.003968	0.003968	0.0064	0.000000
16.5	0.011905	0.007937	-40.5485	-0.003968
17.5	0.019841	0.015873	-22.3130	-0.003968
19.0	0.051587	0.047619	-8.0037	-0.003968
20.5	0.091270	0.075397	-19.1057	-0.015873
22.0	0.142857	0.154762	8.0044	0.011905

CASE 3 tie pattern: (1,3,1,2,1,1,1), sample sizes: (m, n)=(5, 5)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
15.0	0.003968	0.003968	0.0064	0.000000
16.5	0.011905	0.007937	-40.5485	-0.003968
18.0	0.015873	0.015873	0.0001	0.000000
19.0	0.043651	0.047619	8.7007	0.003968
20.0	0.071429	0.075397	5.4061	0.003968
22.0	0.146825	0.154762	5.2646	0.007937

CASE 4 tie pattern: (1,2,1,1,1,1,2,2,1,2), sample sizes: (m, n)=(8, 6)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
36.5	0.000666	0.000333	-69.3146	-0.000333
38.5	0.001665	0.001332	-22.3143	-0.000333
40.0	0.003996	0.003996	0.0001	0.000000
42.5	0.010656	0.009990	-6.4538	-0.000666
45.0	0.027306	0.029637	8.1918	0.002331
48.5	0.073260	0.070929	-3.2334	-0.002331
50.5	0.116883	0.114219	-2.3055	-0.002664

CASE 5 tie pattern: (2,2)×2, sample sizes: (m, n)=(8, 8)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
40.0	0.001088	0.000932	-15.4334	-0.000156
44.0	0.005905	0.005206	-12.6006	-0.000699
50.0	0.035664	0.032479	-9.3559	-0.003185
54.0	0.086713	0.080264	-7.7280	-0.006449
56.0	0.126263	0.117249	-7.4064	-0.009014

pseudo p - values are computed with eq.(3).

Table 2　Discrepancy between exact and pseudo p - values: $21 \leqq m + n \leqq 25$

CASE 6 tie pattern: (3,3,4,3,4,5,3), sample sizes: (m, n)=(15, 10)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
145.0	0.001919	0.002218	14.4649	0.000299
150.5	0.005607	0.005759	2.6759	0.000152
155.5	0.012308	0.013208	7.0578	0.000900
159.0	0.020908	0.023762	12.7965	0.002854
166.0	0.052686	0.057567	8.8596	0.004881
173.0	0.113084	0.118977	5.0796	0.005893

CASE 7 tie pattern: (3,3,4,3,4,5), sample sizes: (m, n)=(12, 10)

Wilcoxon statisitic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
79.0	0.000005	0.000003	-48.0334	-0.000002
95.0	0.001381	0.001718	21.8409	0.000337
106.5	0.016491	0.017914	8.2766	0.001423
110.0	0.030159	0.034575	13.6657	0.004416
115.0	0.063757	0.070119	9.5111	0.006362
120.0	0.121809	0.127151	4.2925	0.005342

CASE 8 tie pattern: (1,2,3,4,5,6), sample sizes: (m, n)=(10, 11)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
65.0	0.000340	0.000394	14.7620	0.000054
75.0	0.004950	0.006359	25.0519	0.001409
78.5	0.011414	0.012078	5.6521	0.000664
83.0	0.026803	0.030500	12.9228	0.003697
86.5	0.046479	0.049309	5.9102	0.002830
89.5	0.073359	0.075868	3.3635	0.002509
95.0	0.146001	0.157186	7.3816	0.011185

CASE 9 tie pattern: (6,5,4,3,2,1), sample sizes: (m, n)=(10, 11)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
67.0	0.000740	0.000760	2.6429	0.000020
75.0	0.005648	0.006359	11.8605	0.000711
78.5	0.010430	0.012078	14.6675	0.001648
83.0	0.026931	0.030500	12.4464	0.003569
86.5	0.046389	0.049309	6.1040	0.002920
89.0	0.067454	0.075868	11.7554	0.008414
93.5	0.124627	0.125602	0.7797	0.000975

CASE 10 tie pattern: (2,2)×6, sample sizes: (m, n)=(12, 12)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
104.0	0.003666	0.003406	-7.3601	-0.000260
110.0	0.010929	0.010245	-6.4669	-0.000684
114.0	0.020541	0.019361	-5.9184	-0.001180
120.0	0.046726	0.044367	-5.1807	-0.002359
124.0	0.075053	0.071584	-4.7325	-0.003469
128.0	0.114275	0.109460	-4.3045	-0.004815

pseudo p - values are computed with eq.(3).

Table 3 Discrepancy between exact and pseudo p - values: $26 \leqq m + n \leqq 30$

CASE 11 tie pattern: (3,3,4,3,4,5,3,5), sample sizes: (m, n)=(15, 15)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
170.0	0.003898	0.004321	10.3001	0.000423
178.0	0.010678	0.011748	9.5504	0.001070
186.0	0.025627	0.027765	8.0123	0.002138
193.0	0.050223	0.053223	5.8015	0.003000
202.0	0.104359	0.108428	3.8252	0.004069

CASE 12 tie pattern: (2,2)×7, sample sizes: (m, n)=(13, 15)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
131.5	0.004074	0.003556	-13.5879	-0.000518
137.5	0.009606	0.008514	-12.0706	-0.001092
145.5	0.025723	0.023197	-10.3366	-0.002526
151.5	0.048483	0.044234	-9.1713	-0.004249
161.5	0.117054	0.108680	-7.4228	-0.008374

CASE 13 tie pattern: (2,2)×7, sample sizes: (m, n)=(12, 16)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
118.0	0.004424	0.004171	-5.8819	-0.000253
124.0	0.010411	0.009874	-5.2983	-0.000537
130.0	0.022065	0.021035	-4.7814	-0.001030
138.0	0.052071	0.049952	-4.1549	-0.002119
146.0	0.106393	0.102655	-3.5764	-0.003738

CASE 14 tie pattern: (3,3,4,3,4,5,4), sample sizes: (m, n)=(12, 14)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
104.0	0.000807	0.001020	23.4249	0.000213
114.0	0.005454	0.006335	14.9716	0.000881
118.0	0.010647	0.011682	9.2760	0.001035
124.5	0.025857	0.026322	1.7812	0.000465
130.0	0.048476	0.053000	8.9223	0.004524
138.0	0.108623	0.115577	6.2056	0.006954

CASE 15 tie pattern: (1,2,3,4,5,6,7), sample sizes: (m, n)=(10, 18)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
89.0	0.002513	0.003066	19.9034	0.000553
99.5	0.012801	0.013627	6.2526	0.000826
105.0	0.025433	0.028632	11.8478	0.003199
111.5	0.052788	0.054592	3.3595	0.001804
118.5	0.101820	0.103926	2.0469	0.002106

CASE 16 tie pattern: (7,6,5,4,3,2,1), sample sizes: (m, n)=(10, 18)

Wilcoxon	exact	pseudo	difference rate	difference
statistic	p - value	p - value	(log_e(pseudo/exact)×100)	(pseudo - exact)
89.5	0.002727	0.003066	11.7309	0.000339
99.5	0.011918	0.013627	13.4000	0.001709
105.0	0.024970	0.028632	13.6850	0.003662
111.5	0.052438	0.054592	4.0248	0.002154
118.5	0.102206	0.103926	1.6685	0.001720

pseudo p - values are computed with eq.(3).

Table 4 Discrepancy between exact and pseudo p - values: $32 \leqq \mathrm{m}+\mathrm{n} \leqq 41$

CASE 17 tie pattern: (3,3,4,3,4,5,5,4,3), sample sizes: (m, n)=(15, 19)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
187.5	0.003914	0.004001	2.1903	0.000087
195.5	0.009211	0.009338	1.3725	0.000127
201.0	0.015590	0.016536	5.8935	0.000946
206.0	0.024254	0.025608	5.4326	0.001354
215.0	0.049829	0.051813	3.9035	0.001984
228.0	0.117665	0.120913	2.7230	0.003248

CASE 18 tie pattern: (2,2)×8, sample sizes: (m, n)=(16, 16)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
192.0	0.003057	0.002909	-4.9574	-0.000148
202.0	0.009886	0.009466	-4.3462	-0.000420
212.0	0.026665	0.025671	-3.7999	-0.000994
222.0	0.061392	0.059401	-3.2960	-0.001991
230.0	0.108039	0.104936	-2.9145	-0.003103

CASE 19 tie pattern: (2,(2,2)×9), sample sizes: (m, n)=(17, 21)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
247.5	0.006726	0.006208	-8.0145	-0.000518
253.5	0.011178	0.010372	-7.4796	-0.000806
263.5	0.024039	0.022490	-6.6617	-0.001549
275.5	0.053273	0.050296	-5.7502	-0.002977
287.5	0.104467	0.099466	-4.9060	-0.005001

CASE 20 tie pattern: (3,3,4,3,4,5,5,4,3,4,3), sample sizes: (m, n)=(20, 21)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
328.0	0.007522	0.007924	5.2015	0.000402
338.0	0.015583	0.016240	4.1284	0.000657
348.0	0.029790	0.030857	3.5193	0.001067
358.0	0.053188	0.054641	2.6959	0.001453
368.0	0.088679	0.090589	2.1315	0.001910
378.0	0.138733	0.141196	1.7597	0.002463

CASE 21 tie pattern: (1,1,1,2,1,1,2,1)x4, sample sizes: (m, n)=(20, 20)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
330.0	0.014910	0.015206	1.9633	0.000296
340.0	0.029272	0.029782	1.7270	0.000510
350.0	0.053229	0.054039	1.5094	0.000810
360.0	0.090129	0.091315	1.3078	0.001186
370.0	0.142772	0.144384	1.1224	0.001612

CASE 22 tie pattern: (6,5,4,3,2,(1,1)×10), sample sizes: (m, n)=(20, 20)

Wilcoxon statistic	exact p - value	pseudo p - value	difference rate (log_e(pseudo/exact)×100)	difference (pseudo - exact)
286.5	0.000263	0.000265	0.8566	0.000002
306.5	0.002124	0.002133	0.4403	0.000009
326.5	0.011352	0.011359	0.0648	0.000007
347.0	0.044349	0.045543	2.6568	0.001194
365.0	0.113630	0.115749	1.8480	0.002119

pseudo p - values are computed with eq.(3).