



## Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What $N$ ? Graphical Insights

Carine A. Bellera  
Institut Bergonié

Marilyse Julien  
James A. Hanley  
McGill University

Copyright © 2010 by Carine A. Bellera, Marilyse Julien, and James A. Hanley all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

---

**Key Words:** Graphics; Mann-Whitney U statistic; Nonparametrics; Normal approximation; Ranking methods; Sampling distribution; Wilcoxon rank-sum statistic; Wilcoxon signed-rank statistic.

### Abstract

The Wilcoxon statistics are usually taught as nonparametric alternatives for the 1- and 2-sample Student- $t$  statistics in situations where the data appear to arise from non-normal distributions, or where sample sizes are so small that we cannot check whether they do. In the past, critical values, based on exact tail areas, were presented in tables, often laid out in a way that saves space but makes them confusing to look up. Recently, a number of textbooks have bypassed the tables altogether, and suggested using normal approximations to these distributions, but these texts are inconsistent as to the sample size  $n$  at which the standard normal distribution becomes more accurate as an approximation. In the context of non-normal data, students can find the use of this approximation confusing. This is unfortunate given that the reasoning behind—and even the derivation of—the exact distributions

can be so easy to teach but also help students understand the logic behind rank tests. This note describes a heuristic approach to the Wilcoxon statistics. Going back to first principles, we represent graphically their exact distributions. To our knowledge (and surprise) these pictorial representations have not been shown earlier. These plots illustrate very well the approximate normality of the statistics with increasing sample sizes, and importantly, their remarkably fast convergence.

## 1. Introduction

Nonparametric statistics are used in data analyses where the assumption of normality does not hold or cannot be ascertained due to limited data. Most of the distribution-free methods employed nowadays were developed in the first half of the 20th Century; they include, among many others, permutation tests by Fisher (1935), run tests by Wald and Wolfowitz (1940), and the Wilcoxon (1945) statistics.

The Wilcoxon signed-rank and the Wilcoxon rank-sum statistics were developed, respectively, as nonparametric alternatives for the 1- and 2-sample Student  $t$ -tests. They can be presented in introductory statistics courses and textbooks; however, the calculation of tail areas can be particularly confusing for new statistics users. First, recursion formulas used to generate the exact distributions are, to our knowledge, never presented. Also, tables of the exact distributions are not systematically provided, and when they are, they are often laid out in a way that saves space but makes them confusing to look up and very difficult for students to use.

The purpose of our note is to present a heuristic approach to how fast the distributions of the Wilcoxon statistics approach normality by relying on graphical displays. To our knowledge, earlier publications considering these distributions focused essentially on numerical issues, while a simple graphical approach seems to have been ignored. Our paper proceeds as follows. We review the Wilcoxon statistics for matched and unmatched data. We describe how some introductory textbooks present these statistics, and in particular what, if anything, they say about the exact and approximate distributions. Since computer packages tend to drive what is taught/learned, we assess how some standard statistical packages handle these statistics, and whether they provide exact computations. Finally, and most importantly, we graphically display the exact distributions of the Wilcoxon statistics based on the original recursion formulas. We emphasize the pertinence and attractiveness of such graphical displays in the teaching of these statistics to new statistics users. Instead of letting students take the asymptotic properties for granted, or suggesting various sample sizes for the normality assumption to hold, we propose using these graphical representations to bolster the recommendations about when one can or cannot use these approximations.

## 2. The Wilcoxon Statistics

In 1945, Wilcoxon introduced the signed-rank and the rank-sum statistics in the context of treatment comparisons (Wilcoxon 1945):

The comparison of two treatments generally falls into one of the following categories: (a) we may have a number of replications for each of the two treatments, which are unpaired, or (b) we may have a number of paired comparisons leading to a series of differences, some of which may be positive and some negative... The object of the present paper is to indicate the possibility of using ranking methods, that is, methods in which scores  $1, 2, 3, \dots, n$  are substituted for the actual numerical data, in order to obtain a rapid approximate idea of the significance of the differences in experiments of this kind.

We start by briefly reviewing these two statistics.

### 2.1 The Wilcoxon Signed-Rank Statistic for Paired Data

The Wilcoxon signed-rank statistic pertains to the comparison of paired data and was illustrated by Wilcoxon (1945) using heights of *Zea mays* (corn) plants. These data were generated by Charles Darwin, and used by Fisher (1935) to illustrate his permutation test. Darwin was interested in the difference in heights between cross- and self-fertilized corn plants of the same pair, planted in the same pot. Denote by  $(X_i, Y_i)$  such pairs of observations, where  $i = 1, \dots, n$ ; in Darwin's experiment,  $n = 15$ . The height difference is measured using a numerical scale, with positive scores indicating the superiority of cross- over self-fertilization. Wilcoxon proposed ranking the  $n$  differences in order of size, disregarding the sign. The Wilcoxon signed-rank statistic,  $W_+$ , is then defined as the sum of the ranks of the positive values, which ranges from 0 (none of the within-pair differences are positive) to  $n(n+1)/2$  (all of the within-pair differences are positive). Similarly, one can compute  $W_-$ , the sum of the ranks of the negative values; clearly, the overall sum  $W_+ + W_- = n(n+1)/2$ .

The null hypothesis of no treatment effect will be rejected if  $W_+$  is more extreme than might be expected to occur by chance alone. Thus, a test based on the Wilcoxon signed-rank statistics consists of evaluating the probability of a value of  $W_+$  equal to or more extreme than that observed (similarly, an equivalent test can be constructed based on  $W_-$ ). The null distribution of  $W_+$ , that is, the probability of  $W_+$  taking on all its possible values assuming randomization, is established as follows. Let  $F_W(w|n)$  denote the number of ways to choose a subset of the integers between one and  $n$  such that their sum is  $w$ , where  $w = 0, 1, \dots, n(n+1)/2$ . Consider the example of  $n = 3$  pairs of observations, with the  $2^3$

possible patterns of signed ranks  $\pm 1, \pm 2, \pm 3$ . If all within-pair differences are positive, that is, if the signed ranks are  $+1, +2$ , and  $+3$ , the sum of the positive ranks is  $W_+ = 6$ . Since there is only one such configuration, then  $F_W(6|3) = 1$ . There is also just  $F_W(5|3) = 1$  way, namely via  $(1-, 2+, 3+)$ , to achieve  $w = 5$ , and similarly there is  $F_W(4|3) = 1$  way to obtain  $w = 4$ . There are  $F_W(3|3) = 2$  configurations where the sum of the positive ranks is 3, that is,  $(+1, +2, -3)$  and  $(-1, -2, +3)$ . There is only one way to obtain  $w = 2, 1$ , and 0, so  $F_W(2|3) = F_W(1|3) = F_W(0|3) = 1$ . Thus, the statistic  $W_+$  is distributed on  $\{0, \dots, 6\}$  with probabilities  $\{\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{2}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\}$ . For larger samples,  $F_W(w|n)$  can be obtained using the following recursive formula (McCornack 1965):

$$F_W(w|n) = F_W(w|n-1) + F_W(w-n|n-1). \quad (2.1)$$

As there are  $2^n$  configurations, the lower tail probability is given by

$$Pr(W_+ \leq w|n) = \frac{\sum_{s=0}^{s=w} F_W(s|n)}{2^n}.$$

Wilcoxon (1945) and Wilcoxon et al. (1963) tabulated critical values for various sample sizes and probability levels up to 50 pairs of observations; extended tables were provided by McCornack (1965). These extensive tabulations were derived long before the use of modern and fast computers, requiring rather tedious work. Although mainframe computers were heavily used in the 1960s, they were on the order of a million times slower than computers today. Finally, the asymptotic properties were also investigated by Wilcoxon (1947) who showed that the distribution tends to a normal one with mean  $n(n+1)/4$  and variance  $n(2n+1)(n+1)/24$ .

In the previous discussion, we focused on distinct differences. In some instances, however, some of the differences may be numerically equal. Differences are then assigned the average of the tied ranks. For example, if differences are  $(10, 12, 12, 15)$ , then the rank of the two middle differences is 2.5. The relationships between probabilities given by Equation (2.1) no longer holds, and the exact critical values derived from untied observations become more conservative. As  $W$  will usually take on less values, the normal approximation will not be as close as with distinct differences. For a discussion on tied differences, we refer the reader to Lehmann (1998).

## 2.2 The Wilcoxon Rank-Sum Statistic for Independent Samples

The Wilcoxon rank-sum statistic pertains to the comparison of two independent samples (Wilcoxon 1945). Assume  $n_1$  and  $n_2$  subjects are randomly assigned to two treatments ( $n_1 \leq n_2$ ) and let  $X_i$  and  $Y_j$  denote the observations for each treatment group ( $i = 1, \dots, n_1$  and  $j = 1, \dots, n_2$ ). The outcomes of the treatment are measured using a numerical scale, with high scores indicating a better outcome. The  $n_1 + n_2$  observations are ranked

according to their scores. For the moment, suppose that the observations are all distinct. If we denote by  $R_i$  the rank of  $X_i$  in the combined sample of  $X$ 's and  $Y$ 's, then the rank-sum statistic is defined as the sum of the ranks in the smaller sample:  $W_1 = R_1 + \cdots + R_{n_1}$ . Similarly, one can define  $W_2$ , the sum of the ranks in the larger sample  $W_2 = R_1 + \cdots + R_{n_2}$ . The smallest value of  $W_1$  occurs when all  $X_1, \dots, X_{n_1}$  are ranked smaller than each of  $Y_1, \dots, Y_{n_2}$ , that is,  $\min[W_1] = n_1(n_1 + 1)/2$  and  $\max[W_2] = n_1n_2 + n_2(n_2 + 1)/2$ . Conversely,  $W_1$  reaches its maximum value when all  $X_1, \dots, X_{n_1}$  are ranked greater than each of  $Y_1, \dots, Y_{n_2}$ , leading to  $W_1 = n_1n_2 + n_1(n_1 + 1)/2$  and  $W_2 = n_2(n_2 + 1)/2$ . The Wilcoxon rank-sum statistics are thus related through the equation  $W_1 + W_2 = \frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1)$ , and tests based on either one are equivalent.

The hypothesis of no treatment effect will be rejected if  $W_1$  is sufficiently extreme. Thus, a test based on the Wilcoxon rank-sum statistic consists of evaluating the probability of observing a value of  $W_1$  equal to or more extreme than the value actually observed. The exact distribution can then be tabulated; however, it requires triple-entry tables, leading to a formidable publication problem given the many pages that would be required. [Mann and Whitney \(1947\)](#) proposed an equivalent statistic which improved this tabulation problem. Instead of considering the sum of the ranks, they considered the  $n_1 \times n_2$  possible pairs  $(X_i, Y_j)$  and introduced the two following statistics:

$$U_{XY} = \text{number of pairs for which } x_i < y_j$$

and

$$U_{YX} = \text{number of pairs for which } x_i > y_j.$$

The Mann-Whitney  $U$ -statistics lie between 0 and  $n_1n_2$  and, in the absence of tied observations, are linked through the relationship  $U_{XY} + U_{YX} = n_1n_2$ . Thus, tests based on either statistic are equivalent. In addition, the Wilcoxon and Mann-Whitney statistics are linked through the following relationships:

$$U_{YX} = W_1 - n_1(n_1 + 1)/2,$$

and

$$U_{XY} = W_2 - n_2(n_2 + 1)/2.$$

Therefore, the probability of observing a value of  $W_1$  greater than or equal to that observed equals the probability of a value of  $U_{YX}$  greater or equal to that observed. Tests based on the Wilcoxon rank-sum and Mann-Whitney statistics are thus equivalent, and since  $U_{YX}$  and  $U_{XY}$  have the same null distribution, tabulation of the  $U$ -statistics is considerably reduced compared to that of the Wilcoxon statistics. The exact distribution of the  $U$ -statistic can be derived through recursion. Let  $F_U(u|n_1, n_2)$  denote the frequency (regardless of the order)

with which a  $Y$  precedes an  $X$   $u$  times in samples of size  $n_1 + n_2$ , then:

$$F_U(u|n_1, n_2) = F_U(u - n_2|n_1 - 1, n_2) + F_U(u|n_1, n_2 - 1), \quad (2.2)$$

with the initial conditions  $F_U(u|n_1, n_2) = 0$  if  $u < 0$ ,  $n_1 < 0$ ,  $n_2 < 0$  or  $u > n_1 n_2$ , and  $F_U(0|n_1, n_2) = 1$  if  $n_1 \geq 0$  and  $n_2 \geq 0$ .

Under randomization, there are  $(n_1 + n_2)!/(n_1!n_2!)$  equally likely possible rankings of the combined sample. Thus the probability distribution function is given by:

$$Pr(U = u|n_1, n_2) = \frac{n_1!n_2!}{(n_1 + n_2)!} F_U(u|n_1, n_2).$$

For large sample sizes, [Mann and Whitney \(1947\)](#) proved that the statistics are normally distributed with mean  $n_1 n_2 / 2$  and variance  $n_1 n_2 (n_1 + n_2 + 1) / 12$ . [Wilcoxon \(1945\)](#) tabulated critical values for the special case of samples of equal sizes (i.e., for  $n_1 \leq n_2 \leq 8$ ). [Mann and Whitney \(1947\)](#) investigated the case of unequal sample sizes up to a maximum sample size of 8. Later, critical values were tabulated for multiple significance levels (down to 0.0005) and sample sizes (up to 50 observations) ([Milton 1964](#); [Verdooren 1963](#); [Wilcoxon et al. 1970](#)).

The relationship between the Wilcoxon rank-sum and the Mann-Whitney statistics is just a shift in location. Thus, for the Wilcoxon statistics as such, the recursion formula can be easily derived (for details see the [Appendix](#)). Let  $F_W(w|n_1, n_2)$  denote the frequency, regardless of the order, with which the sum of the positive ranks in the smallest sample equals  $w$  in samples of size  $n_1$  and  $n_2$ , then:

$$F_W(w|n_1, n_2) = F_W(w - n_1 - n_2|n_1 - 1, n_2) + F_W(w|n_1, n_2 - 1), \quad (2.3)$$

with the initial conditions  $F_W(w|n_1, n_2) = 0$  if  $w < \frac{n_1(n_1+1)}{2}$ ,  $n_1 < 0$ ,  $n_2 < 0$  or  $w > n_1 n_2 + \frac{n_1(n_1+1)}{2}$ , and  $F_U(\frac{n_1(n_1+1)}{2}|n_1, n_2) = 1$  if  $n_1 \geq 0$  and  $n_2 \geq 0$ .

Again, we have presented the rank-sum statistics in the context of distinct observations. Tied observations are handled in the same way as for the signed-rank statistics except that average ranks are used for all tied observations. Similarly, relationships provided by Equations (2.2) and (2.3) no longer hold. The exact critical values derived will be more conservative, and the normal approximation will not be as close as with distinct observations. For more information, see [Lehmann \(1998\)](#).

We close this section with an interesting historical note. Although nowadays we usually refer to the work by Wilcoxon or Mann and Whitney, tests based on similar statistics were proposed much earlier. In an interesting review, [Kruskal \(1957\)](#) found that seven independent versions of the Wilcoxon test for unpaired data, or similar procedures, had been proposed earlier—some as early as 1914.

### 3. Teaching Nonparametrics in Introductory Statistics Textbooks and Software Packages

#### 3.1 Textbooks

We examined a convenience sample of 12 introductory statistics textbooks and three non-parametric statistics textbooks available to students at the McGill University Science Library. None of these books included graphical displays of the distributions of the Wilcoxon statistics. Only one book (Lehmann 1998) included recursive formulas.

#### 3.2 Software

Similarly, we investigated to what extent exact computations were used when tests based on the Wilcoxon statistics are requested. We focused on seven commonly used statistics packages: MINITAB (Minitab Inc., State College, PA, USA), STATA (Stata Corporation, College Station, TX, version 9.2), SAS (SAS Institute Inc., Cary, NC, USA, version 9.1), SPSS (SPSS Inc., Chicago, version 16.0), Splus (Mathsoft Inc., Seattle, version 6.2), R (The R Foundation for Statistical Computing, TX, Version 2.7), and StatXact (Cytel Inc. Cambridge, MA, version 8).

With regard to the test based on the signed-rank statistics, MINITAB provides only normal approximations (`1-sample Wilcoxon` command). The documentation provided by the STATA package (`signrank` command) does not indicate the algorithm used for computations. The Z-statistic provided in the output suggested that results are based on the normal approximation, regardless of the sample size. The SPSS documentation indicates that the  $p$ -values are based on a normal approximation (`npar` command, `paired` option). The SAS documentation specifies that exact computations are applied for 20 pairs of observations or fewer, and an approximation to the  $t$ -distribution is used otherwise (`univariate` procedure, `Output` statement, `probs` option). StatXact, a statistical package especially devoted to exact nonparametric inference, provides exact computations (`Wilcoxon Signed-rank` command, `Exact test method` option). As detailed in the help documentation, both R and Splus (`stats` library) allow the user to specify whether exact or approximate computations are required (`wilcox.test(paired=T)`, `exact` option). By default, Splus provides exact tests for samples with less than 25 pairs of observations. If there are ties or more than 25 pairs of observations, only the normal approximation is available. Similarly, exact results are available in R when there are 50 pairs of observations or fewer. Interestingly, the documentation for the `pwilcox` and `dwilcox` functions in R provide examples on how to graph the exact distribution and density functions.

With regard to Wilcoxon rank-sum statistics, [Bergmann et al. \(2000\)](#) investigated outcomes of the rank-sum statistics from different statistics packages. With respect to the use of exact computations, our current observations suggest that there have not been any changes since then. MINITAB only provides normal approximations (Mann-Whitney command). The STATA documentation is unclear. However, the outputs suggest that the normal approximation is applied whatever the sizes of the samples (ranksum command). SPSS provides exact computations for 40 observations or less, otherwise a normal approximation is performed (npar command, M=W subcommand). By default, SAS provides normal approximations even for small sample sizes (procedure npar1way) but the user can specify if exact computations are required (exact statement). Exact computations are available from StatXact (Wilcoxon-Mann-Whitney command, Exact test method option). Both R and Splus (stats library) provide rather extensive information. In the absence of ties, and when the two datasets have 49 or fewer observations, exact computations are provided by default (Wilcox.test(paired=F) command); otherwise, a normal approximation is applied. Note however, that in the absence of ties, the user can request exact computations, whatever the sample sizes (exact option). Again, the documentation for the psignrank and dsignrank functions in R provide examples on how to graph the exact distribution and density functions.

Aside from StatXact, which is specifically developed for nonparametric inference, the other statistical packages show important variations in the amount and quality of the information provided by their documentation and outputs. While some software packages are flexible and allow the users to specify if exact results are needed, others provide tests based on approximate tail areas only. The existence of statistical software especially devoted to nonparametric inference might explain why nonparametric methods presented in most general statistical software are limited to results based on normal approximations. However, given that recursion is easy and fast, we feel that exact methods for nonparametric statistics could be easily implemented in most statistical software.

## 4. Exact Distribution of the Wilcoxon Statistics

### 4.1 Numerical Approach

Critical values for the Wilcoxon statistics have been tabulated extensively. [Wilcoxon \(1947\)](#) tabulated critical values for up to 10 pairs of observations for the signed-rank statistics and up to 8 observations per sample for the rank-sum statistics (he only considered samples of equal sizes for the case of unpaired data). Recognizing that providing such tables soon “*becomes impractical with larger numbers of replicates*” ([Wilcoxon 1947](#)), he investigated the asymptotic properties of these statistics. He then provided tables based on normal ap-

proximation starting with as few as six pairs of observations for the signed-rank statistics, and five observations per sample for the rank-sum statistics. Similarly, Mann and Whitney (1947) considered the case of unpaired data with samples of unequal sizes and reported exact tables only up to  $n_1 = n_2 = 8$  and concluded that “*at this point the distribution is almost normal*” (Mann and Whitney 1947). Very early, it was thus acknowledged that the normal approximation was acceptable for rather small sample sizes.

Afterwards, however, additional tables of critical values were provided for larger sample sizes and smaller significance levels for both paired (Wilcoxon 1947; McCornack 1965; Lehmann 1998) and unpaired data (Mann and Whitney 1947; Fix and Hodges 1955; Jacobson 1963; Verdooren 1963; Milton 1964; Lehmann 1998). Statisticians also focused on numerical evaluations of the asymptotic properties of the signed-rank (Wilcoxon 1947; Fellingham and Stocker 1964; McCornack 1965; Claypool and Holbert 1974) and rank-sum statistics (Wilcoxon 1947; Buckle et al. 1969; Di Bucchianico 1999; Bean et al. 2004). Critical values of the Wilcoxon statistics were computed and compared numerically applying both the exact method or approximations (normal in most instances, but saddle-point, Edgeworth or uniform approximations have also been investigated).

## 4.2 Graphical Approach

Surprisingly, while there have been several meticulous numerical evaluations of the asymptotic properties of the Wilcoxon statistics, none of the textbooks, published papers, and very few software that we surveyed proposed graphical comparisons. When plots were provided, those usually depicted the percent (relative or absolute) error as a function of the probability level. While one can easily find plots of the binomial, Poisson, or other distributions, it seems that the Wilcoxon statistics have been completely ignored. This is both surprising and unfortunate as such plots can be easily generated from first principles using recursive formulas.

***The Wilcoxon signed-rank statistics.*** The exact distribution of the Wilcoxon signed-rank statistics can be easily and nicely illustrated using the recursion formula represented by Equation (2.1). We plotted the frequencies of the sum of the positive ranks and the sum of the negative ranks for samples of size  $n = 1$  to 7 using a representation similar to that of Pascal’s triangle. In Figure 1, the sums of the positive ( $W_+$ ) and negative ranks ( $W_-$ ) are given respectively on the x- and y-axis. Each diagonal corresponds to the distribution of the signed-rank statistic for one particular sample size (from  $n = 1$  in the lower-left corner to  $n = 7$  in the upper-right corner). For example, the diagonal with  $n = 7$  corresponds to the frequencies of all the possible combinations ( $W_+, W_-$ ) when there are seven pairs of observations. In this particular case, there are five possibilities of obtaining the pair

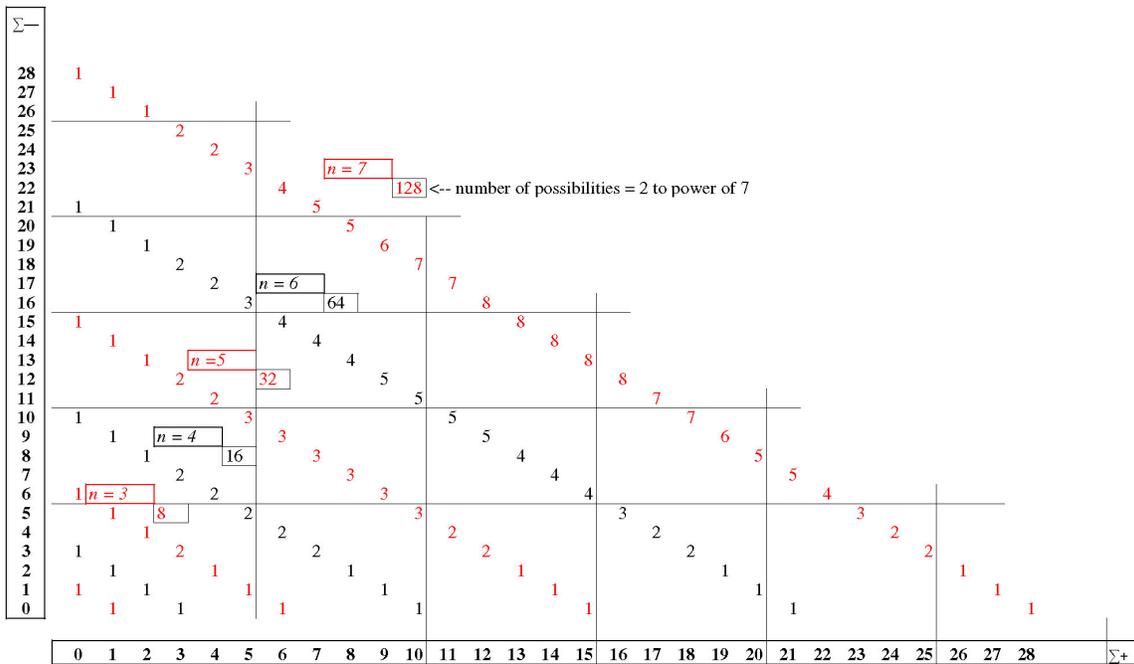


Figure 1. Distribution of the Wilcoxon signed-rank statistic, for samples (i.e., diagonal) of sizes  $n = 1$  to  $n = 7$ .

$(W_+, W_-) = (8; 20)$ , that is, using notation of Equation (2.1),  $F(8|7) = 5$ . This entry of 5 is obtained as the sum of the 1 immediately to its left and the 4 directly below it, equivalent to applying the recursion formula :  $F(8|7) = F(8|6) + F(1|6) = 4 + 1 = 5$ . This enumeration illustrates well the symmetry of the signed-rank statistic, as well as its closeness to the normal distribution as the sample size increases.

We also plotted together the exact distribution with the normal curve, and both tend to overlap relatively quickly (Figure 2). This is in accordance with Wilcoxon’s own suggestion that as few as six pairs of observations are necessary for the normal approximation to hold.

**The Wilcoxon rank-sum statistics.** Similarly, we focused on the graphical display of the distribution of the Wilcoxon rank-sum statistic and have represented all the equally likely configurations of the sum of the ranks in the smaller of 2 samples (Figure 3). All the possible individual ranks are shown, and all those with the same sum are represented one above the other. For illustration, consider the bottom of Figure 3, where we have assumed two independent samples of sizes  $n_1 = 3$  and  $n_2 = 5$  are available (i.e.,  $X_1, X_2, X_3, Y_1, Y_2, Y_3, Y_4, Y_5$ ). Each observation can be assigned a different rank between 1 and 8. If the three

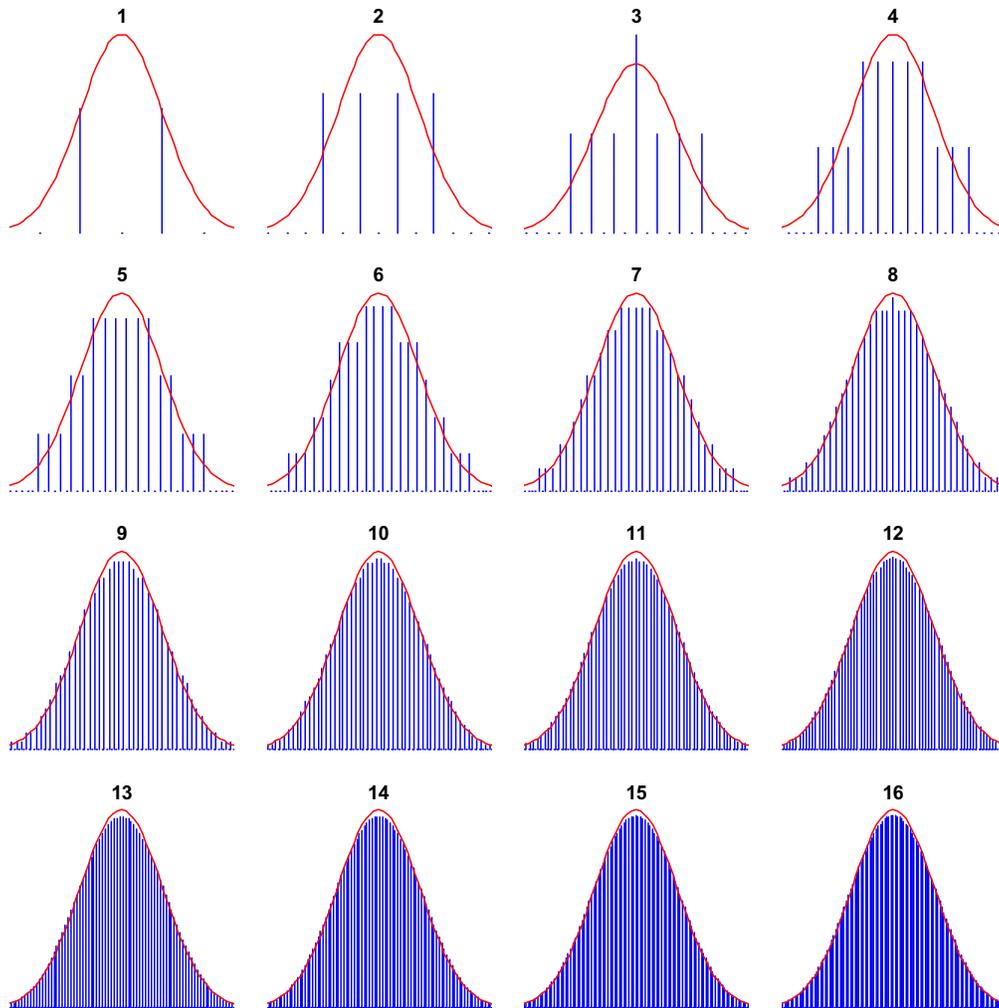


Figure 2. Exact distribution of the Wilcoxon signed-rank statistic for sample of sizes ( $1 \leq n \leq 16$ ), and curve of the normal distribution with mean  $n(n+1)/4$  and variance  $n(n+1)(2n+1)/24$ .

smallest observations are all in the smaller sample, then these observations are assigned the ranks  $(1, 2, 3)$ , that is, the sum of the ranks in the smaller sample is  $1 + 2 + 3 = 6$ . Similarly, if the first, second, and fourth observations are in the smaller sample, the sum of the ranks in this sample will be  $1 + 2 + 4 = 7$ . There are, however, two possible configurations that can lead to a sum of the ranks in the smaller sample equal to 8: if the first, second, and fifth smallest observations or if the first, third, and fourth smallest observations are in the smaller sample. In total, there are therefore 56 equally likely configurations of 3 ranks from these 8 observations, starting on the left with ranks  $(1, 2, 3)$  (if all  $X$ 's are smaller than all  $Y$ 's) to the opposite  $(6, 7, 8)$  on the extreme right. Thus, the sum of the 3 ranks ranges from  $1 + 2 + 3 = 6$  to  $6 + 7 + 8 = 21$ . There are, for example, six configurations leading to a

		Sum of Ranks in smaller sample																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1																					
1	2	1																				
1	3	1	2	3	4																	
2	2		1,2	1,3	1,4	2,4	3,4															
2	3		1,2	1,3	1,4	1,5	2,5	3,5	4,5													
2	4		1,2	1,3	1,4	1,5	1,6	2,6	3,6	4,6	5,6											
2	5		1,2	1,3	1,4	1,5	1,6	1,7	2,7	3,7	4,7	5,7	6,7									
2	6		1,2	1,3	1,4	1,5	1,6	1,7	1,8	2,8	3,8	4,8	5,8	6,8	7,8							
3	3				1,2,3	1,2,4	1,2,5	1,2,6	1,3,6	1,4,6	1,5,6	2,5,6	3,5,6	4,5,6								
3	4				1,2,3	1,2,4	1,2,5	1,2,6	1,2,7	1,3,7	1,4,7	1,5,7	1,6,7	2,6,7	3,6,7	4,6,7	5,6,7					
3	5				1,2,3	1,2,4	1,2,5	1,2,6	1,2,7	1,2,8	1,3,8	1,4,8	1,5,8	1,6,8	1,7,8	2,7,8	3,7,8	4,7,8	5,7,8	6,7,8		

Figure 3. Distribution of the Wilcoxon rank-sum statistic, for samples of sizes  $n_1 = 1$  to 3 and  $n_2 = 1$  to 5.

sum of the ranks equal to 14, that is,  $F_W(14|3,5) = 6$ . Applying recursion formula (2.3):  $F_W(14|3,5) = F_W(6|2,5) + F_W(14|3,4) = 2 + 4$  as illustrated.

We have plotted the distribution of the Wilcoxon rank-sum statistic,  $W_1$ , for cases up to 8 observations in the largest group (samples of sizes  $n_1 = 1$  to 5 and  $n_2 = 1$  to 8), as shown in Figure 4. For each scenario, we superimposed the curve of the normal distribution. As suggested earlier (Mann and Whitney 1947), the normal approximation appears appropriate very quickly.

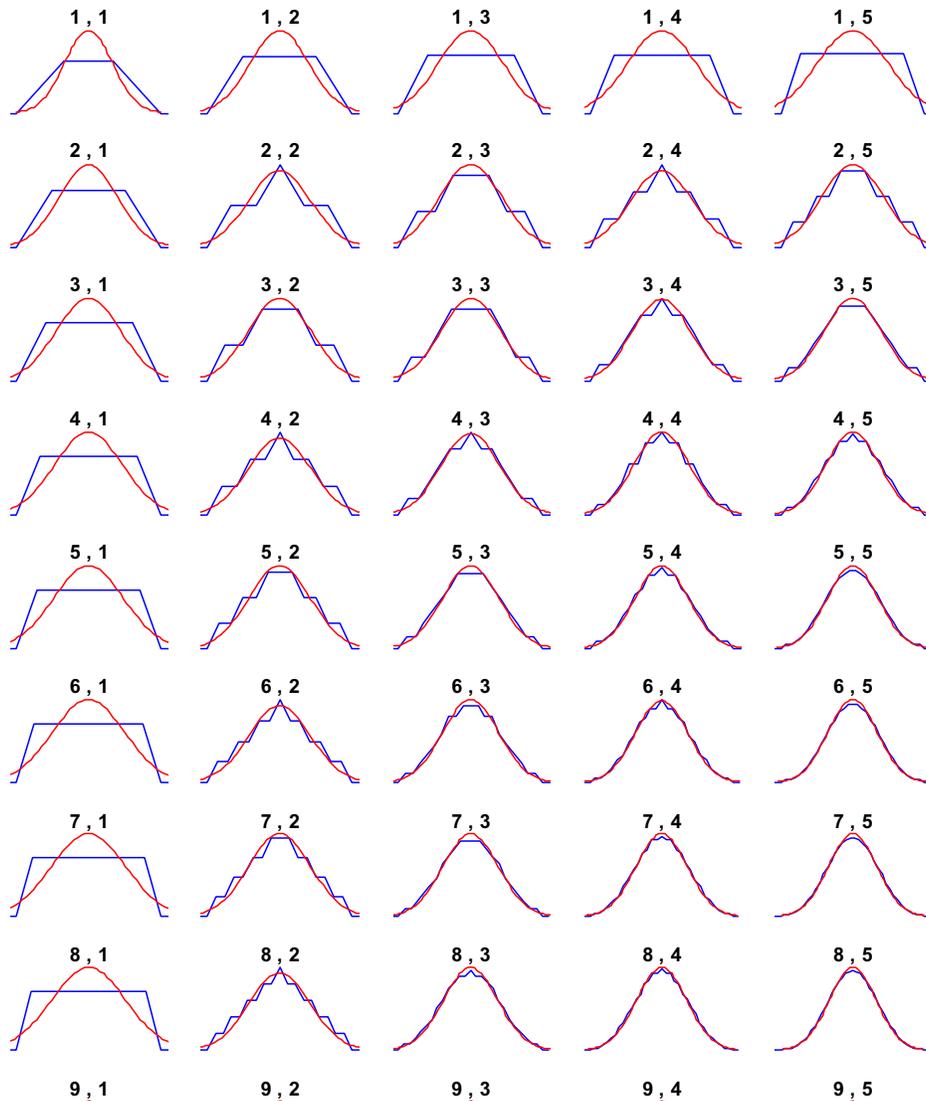


Figure 4. Exact distribution of the Wilcoxon rank-sum statistic for samples of sizes  $n_1 = 1$  to 5 and  $n_2 = 1$  to 8, and curve of the normal distribution with mean  $(n_1 + n_2 + 1)/2$  and variance  $n_1 n_2 (n_1 + n_2 + 1)/12$ .

Figure 1 and Figure 3 were obtained using Excel. Figure 2 and Figure 4 were graphed in R; given R functions are available (see [software section](#)), obtaining these figures did not require extra programming.

## 5. Conclusion

We have provided a heuristic approach to the Wilcoxon statistics for both paired and unpaired observations, based on simple graphical representations of their exact distributions. These plots show that the normal approximation can be appropriate for relatively small

sample sizes, as originally suggested by Wilcoxon as well as Mann and Whitney. Extensive tabulations of the exact Wilcoxon distribution have been produced going in some instances up to 50 observations, and down to small significance levels. Such tables however, can be particularly complex to look up for new statistics users and may not be the best way to introduce the asymptotic properties of the Wilcoxon statistics.

While writing this note, we surveyed multiple publications and textbooks and were surprised to see that the Wilcoxon statistics had never been illustrated in their simplest form. Indeed, the only one to suggest plotting them was the author who wrote the documentation for the `psignrank` function (for the probability distribution of the signed-rank statistic) and `pwilcox` functions (for the probability distribution of the rank-sum statistic) functions in the R package. In the dozen introductory textbooks available to us, simple histograms of distributions seem to have been completely ignored. We feel, however, that such representations could be valuable tools when introducing these statistics. In particular, these graphs illustrate very well the asymptotic property of these statistics and could therefore reduce the confusion about using normal distributions for statistics based on non-normal data.

## Appendix: Recursion Formula to Generate Distribution of Wilcoxon Rank-Sum Statistic

Recall that  $F_U(u|n_1, n_2)$  denotes the frequency (regardless of the order) with which a  $Y$  precedes an  $X$   $u$  times in samples of size  $n_1 + n_2$  and can be derived using Equation (2.2). Similarly,  $F_W(w|n_1, n_2)$  denotes the frequency (regardless of the order) with which the sum of the positive ranks in the smallest sample equals  $w$  in samples of size  $n_1 + n_2$ , and recall that the statistics  $U$  and  $W$  are related though  $U_{YX} = W_1 - n_1(n_1 + 1)/2$ . Thus, the left hand-side of Equation (2.2) can be written as:

$$F_U(u|n_1, n_2) = F_W\left(u + \frac{n_1(n_1 + 1)}{2} | n_1, n_2\right)$$

Similarly, for the first term in the right-hand side of Equation (2.2), we have:

$$\begin{aligned} F_U(u - n_2 | n_1 - 1, n_2) &= F_U\left(w - \frac{n_1(n_1 + 1)}{2} - n_2 | n_1 - 1, n_2\right) \\ &= F_W\left(w - n_1\left(\frac{n_1 + 1}{2} - n_2 + \frac{(n_1 - 1)n_1}{2}\right) | n_1 - 1, n_2\right) \\ &= F_W(w - n_1 - n_2 | n_1 - 1, n_2). \end{aligned}$$

Finally, the second term in the right-hand side is given by:

$$F_U(u | n_1, n_2 - 1) = F_W(w | n_1, n_2 - 1)$$

Combining the previous equalities leads to Equation (2.3).

## Acknowledgments

This work was partially supported by grants from The Natural Sciences and Engineering Research Council of Canada (J. H.) and Le Fonds Québécois de la Recherche sur la Nature et les Technologies (J. H.).

---

## References

- Bean, R., Froda, S., and Van Eeden, C. (2004), “The Normal, Edgeworth, Saddlepoint and Uniform Approximations to the Wilcoxon-Mann-Whitney Null-distribution: a Numerical Comparison,” *Nonparametric Statistics*, 16, 279–288.
- Bergmann, R., Ludbrook, J., and Spooren, W. (2000), “Different Outcomes of the Wilcoxon-Mann-Whitney Test from Different Statistics Packages,” *The American Statistician*, 54, 72–77.
- Buckle, N., Kraft, C., and Van Eeden, C. (1969), “An Approximation to the Wilcoxon-Mann-Whitney Distribution,” *Journal of the American Statistical Association*, 64, 591–599.
- Claypool, P.L., and Holbert, D. (1974), “Accuracy of the Normal and Edgeworth Approximations to the Wilcoxon Signed Rank Statistics,” *Journal of the American Statistical Association*, 69, 255–258.
- Di Bucchianico, A. (1999), “Combinatorics, Computer Algebra and the Wilcoxon-Mann-Whitney Test,” *Journal of Statistical Planning and Inference*, 79(2): 349–364.
- Fellingham, S.A., and Stocker, D.J. (1964), “An Approximation for the Exact Distribution of the Wilcoxon Test for Symmetry,” *Journal of the American Statistical Association*, 59, 899–905.
- Fisher, R.A. (1935), *The Design of Experiments*: Oliver & Boyd, Ltd.
- Fix, E., and Hodges, J.L. (1955), “Significance Probabilities of the Wilcoxon Test,” *Annals of Mathematical Statistics*, 26, 301–312.

Hollander, M., and Wolfe, D. (1999), *Nonparametric Statistical Methods* (2nd Ed.), New York: Wiley.

Jacobson, J. (1963), "The Wilcoxon Two-sample Statistic: Tables and Bibliography," *Journal of the American Statistical Association*, 58, 1086–1103.

Kruskal, W.H. (1957), "Historical Notes on the Wilcoxon Unpaired Two-sample Test," *Journal of the American Statistical Association*, 52, 356–360.

Lehmann, E. (1998), *Nonparametrics—Statistical Methods Based on Ranks* (Revised First Edition), San Francisco: Holden-Day Inc.

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

McCornack, R.L. (1965), "Extended Tables of the Wilcoxon Matched Pair Signed Rank Statistic," *Journal of the American Statistical Association*, 60, 864–871.

Milton, R.C. (1964), "An Extended Table of Critical Values for the Mann-Whitney (Wilcoxon) Two-sample Statistic," *Journal of the American Statistical Association*, 59, 925–934.

Siegel, S., and Castellan, N.J. (1988), *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.), New York: McGraw-Hill, Inc.

Verdooren, L.R. (1963), "Extended Tables for Critical Values for Wilcoxon's Test Statistic," *Biometrika*, 50, 177–186.

Wald, A., and Wolfowitz, J. (1940), "On a Test Whether two Samples are from the Same Population," *Annals of Mathematical Statistics*, 11, 147–162.

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

Wilcoxon, F. (1947), "Probability Tables for Individual Comparisons by Ranking Methods," *Biometrics*, 3, 119–122.

Wilcoxon, F., Katti, S.K., and Wilcox R. (1963), *Critical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Rank Test*, Pearl River, N.Y.: American Cyanamid Co. and Florida State University.

Wilcoxon, F., Katti, S.K., and Wilcox R. (1970), "Critical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Rank Test," in *Selected Tables in Mathematical Statistics* ( vol. 1), pp, 171–259, ed. Harter and Owen, Chicago: Makham Publishing Co.

---

Carine A. Bellera

[bellera@bergonie.org](mailto:bellera@bergonie.org)

Department of Clinical Epidemiology and Clinical Research

Institut Bergonié

Regional Comprehensive Cancer Center

Bordeaux, FRANCE

Marilyse Julien

Department of Mathematics and Statistics

McGill University, 805 Sherbrooke Street West

Montreal, Quebec, H3A 2K6

CANADA

James A. Hanley

Department of Epidemiology

Biostatistics, and Occupational Health

Montreal, CANADA

---

[Volume 18 \(2010\)](#) | [Archive](#) | [Index](#) | [Data Archive](#) | [Resources](#) | [Editorial Board](#) |

[Guidelines for Authors](#) | [Guidelines for Data Contributors](#) |

[Guidelines for Readers/Data Users](#) | [Home Page](#) |

[Contact JSE](#) | [ASA Publications](#)