

# 4.1 Wilcoxon signed-rank test

## Overview

The Wilcoxon signed-rank test was described by [Wilcoxon \(1945\)](#). It is designed for the situation where there are **two groups** of values, and any individual value in one group is **paired** with a specific value in the other group. For example, you might have a treatment and a control value for a given response variable across each of a number of different trials. Interest lies in making a formal comparison of treatment vs control values, acknowledging the inherent non-independence of the paired values within each trial. This test is a non-parametric analogue to a classical paired  $t$ -test, and may be implemented in PRIMER under either a directional (one-tailed) or non-directional (two-tailed) alternative hypothesis.

## The null hypothesis

The essential null hypothesis tested here is  $H_0$ : *the distribution of paired differences is stochastically symmetric about zero*. In other words, the rank order of the values within each pair is arbitrary, so the two observations within any pair are exchangeable with one another. We might consider writing this as  $H_0$ : the median of the distribution of paired differences is equal to zero.

Our alternative hypothesis may be *non-directional* and simply assert that the paired differences are stochastically symmetric about some other value that is *not zero*. This might be phrased as  $H_A$ : the median of the distribution of paired differences is not equal to zero. In that case, we have a two-tailed test.

We may, however, assert a *directional* alternative hypothesis, yielding a one-tailed test. For example:

- $H_A$ : The distribution of paired differences is symmetric about some value **greater** than zero (e.g., the median of the distribution of paired differences is positive); or
- $H_A$ : The distribution of paired differences is symmetric about some value **less** than zero (e.g., the median of the distribution of paired differences is negative).

## Description of the test

Consider a set of  $N$  sampling units. For each sampling unit  $i = 1, \dots, N$ , two paired (or matched) observation values have been recorded:  $(x_i, y_i)$ . For any pair  $i$ , let the difference in these paired values be  $d_i = (x_i - y_i)$ . Next, let  $r_i$  be the rank of the absolute value of the differences  $|d_i|$ , with the smallest absolute difference being given a rank of 1 and the largest absolute difference being given a rank of  $N$ . Let  $\text{sgn}(\cdot)$  be a function to attribute an indicative sign such that  $\text{sgn}(d_i) = 1$  if  $d_i > 0$  and  $\text{sgn}(d_i) = -1$  if  $d_i < 0$ .

We can obtain the *signed ranks* as:  $r_i^{\text{sgn}} = \text{sgn}(d_i) \cdot r_i$ .

We define the test statistic,  $W$ , as the sum of all the signed ranks, i.e.:  $W = \sum_{i=1}^N \text{sgn}(d_i) \cdot r_i$

Having obtained an observed value of the test statistic from the data,  $W_{\text{obs}}$ , then (as in many other PRIMER routines), we can obtain a  $p$ -value empirically using an appropriate permutation algorithm. Specifically, under the assumption of exchangeability, we can generate a plausible value of the test-statistic  $W$  under a true null hypothesis by randomizing the ordering of the paired values  $(x_i, y_i)$  separately, within each pair, for each and every sampling unit  $i = 1, \dots, n$ . Once this randomization has been done, we can re-calculate the differences,  $d_i^{\pi}$ , and their associated (unsigned) ranks,  $r_i^{\pi}$ , under permutation, to yield:

$$W^{\pi} = \sum_{i=1}^N \text{sgn}(d_i^{\pi}) \cdot r_i^{\pi}$$

We repeat the above randomization and re-calculation procedure a large number of times (e.g., say  $n_{\text{perm}} = 9999$ ) to obtain a large number of values of  $W^{\pi}$  under a true null hypothesis. The probability ( $p$ -value) associated with the null hypothesis (and two-tailed alternative hypothesis) is then estimated empirically as the proportion of values of  $W^{\pi}$  that are equal to or more extreme (in absolute value) than the observed value of the test-statistic,  $W_{\text{obs}}$ . Thus, letting  $W_k^{\pi}$  be the value of  $W^{\pi}$  obtained for the  $k$ th permutation ( $k = 1, \dots, n_{\text{perm}}$ ), the  $p$ -value is:

$$P = \frac{\sum_{k=1}^{n_{\text{perm}}} (\text{I}(|W_k^{\pi}| \geq |W_{\text{obs}}|) + 1)}{(n_{\text{perm}} + 1)}$$

with  $\text{I}(\text{expression}) = 1$  if  $\text{expression}$  is true and zero otherwise. Note that the '+1' in the numerator and denominator of this fraction is there to acknowledge the inclusion of the observed value as a member of the distribution of  $W$  under a true null hypothesis.

## One-tailed alternative hypotheses

As noted above, we may postulate a more specific alternative hypothesis. In such cases, the test-statistic and randomization procedure are all done the same way as described above, but the  $p$ -value is calculated differently. For example, if our alternative hypothesis is that the median of the distribution of paired differences is *greater than zero*, then the  $p$ -value is calculated as:

$$P = \frac{\sum_{k=1}^{n_{\text{perm}}} (\text{I}(W_k^{\pi} \geq W_{\text{obs}}) + 1)}{(n_{\text{perm}} + 1)}$$

which tallies only the values of  $W_k^{\pi}$  that equal or positively exceed  $W_{\text{obs}}$ , in the right-hand tail of the permutation distribution.

If, on the other hand, our alternative hypothesis is that the median of the distribution of paired differences is *less than zero*, the  $p$ -value is calculated as:

$$P = \frac{\sum_{k=1}^{n_{\text{perm}}} (\text{I}(W_k^{\pi} \leq W_{\text{obs}}) + 1)}{(n_{\text{perm}} + 1)}$$

which tallies only the values of  $W_k^{\pi}$  that equal or negatively exceed  $W_{\text{obs}}$ , in the left-hand tail of the permutation distribution.

## Treatment of tied ranks

What happens to the rank-order values,  $r_i$ , in the event of a *tie*? Let's suppose  $|d_1|$  is the smallest absolute difference (so is given a rank of  $r_1 = 1$ ), that  $|d_2|$  is the second-smallest absolute difference (hence  $r_2 = 2$ ), but then we find that a third value for the difference  $|d_3|$  is precisely (within double precision) equal to  $|d_2|$ ? In other words, suppose  $|d_2|$  and  $|d_3|$  are tied for second place in the ranking of values from smallest to largest.

In this case PRIMER takes a fairly standard approach and *averages* the ranks of tied values. We simply order the values from smallest to largest and give them 'raw' ranks that correspond to the set of ordered integers, i.e., from  $1$  to  $n$ . Then, we replace the ordered integers for any tied values with the average of those integers. Thus, in this case, the ordered integers are  $\{1, 2, 3, \dots\}$ . We have  $r_1 = 1$ , but then both  $r_2$  and  $r_3$  would be given the average of the ordered integers in their place, i.e.,  $r_2 = r_3 = (2+3)/2 = 2.5$ . So, the set of ranks used for the analysis will be  $\{1, 2.5, 2.5, \dots\}$ . The next largest absolute difference would be given the (unsigned) rank value of  $4$  (presuming it is not tied), and so on. Any other tied values are treated in the same way, by averaging their corresponding ordered integer values, and all subsequent calculations to calculate the test statistic, etc., simply carry on from there precisely as described above.

An important point here about the PRIMER implementation of the Wilcoxon signed-rank test is that, given that  $p$ -values are calculated using an appropriate randomization algorithm under a true null hypothesis of exchangeability, these non-parametric tests are indeed exact tests, even in the event of there being ties in the ranks. This contrasts with other available software implementations of the Wilcoxon test (e.g., such as `wilcox.test()` in R), which do not compute exact  $p$ -values if there are tied ranks.

## Treatment of tied pairs (difference = 0)

It is possible to obtain paired values that are identical to one another; i.e.,  $x_i = y_i$ , so  $d_i = 0$ . This is problematic for the test, not because of the ranking procedure (it would clearly be of very low rank, given its small absolute value), but because, for this sampling unit, there is therefore *no sign to attribute to it*. The simplest solution is to omit differences equal to zero from the calculation ( [Wilcoxon \(1949\)](#) ). This is the approach that is implemented in PRIMER.

It might be argued, however, that this withdraws certain evidence in favour of the null hypothesis. An alternative approach, suggested by [Pratt \(1959\)](#) , would be to include the zeros when ranking the absolute differences, but subsequently to exclude them from the calculation of the test-statistic; i.e., to thereafter assert that  $\text{sgn}(0) = 0$ .

From a practical point of view, neither the [Pratt \(1959\)](#) , nor the [Wilcoxon \(1949\)](#) method was found to be universally most efficient ( [Conover \(1972\)](#) ). Also, as PRIMER does not rely on any formal derivation of the distribution of the  $W$  test statistic, but rather uses permutation algorithms to estimate a  $p$ -value empirically, it would seem that the choice between these two

options would have little or no substantive effect on the outcome of PRIMER's implementation of the Wilcoxon test, although this has not been investigated in detail.

## Original test-statistic

We note here, in passing, that the original Wilcoxon test-statistic was defined somewhat differently from the above description, which uses  $W$ . First, let  $R^+$  be the sum of the positively signed ranks, i.e.  $R^+ = \sum_{i=1}^N r_i^{\text{sgn}} \cdot \text{I}(d_i > 0)$  and let  $R^-$  be the sum of the negatively signed ranks, i.e.,  $R^- = \sum_{i=1}^N r_i^{\text{sgn}} \cdot \text{I}(d_i < 0)$ . Wilcoxon's original test-statistic is then given as the minimum of the absolute values of these two quantities; that is,  $T = \min(|R^+|, |R^-|)$ .

The PRIMER implementation provides both  $W$  and  $T$  in the output file for this test. Note that the calculation of a  $p$ -value using an appropriate permutation algorithm under a true null hypothesis of exchangeability ensures that PRIMER achieves an exact test, no matter which test statistic you prefer to report from the output provided.

---

Revision #147

Created 19 November 2024 03:02:44 by Marti

Updated 25 June 2025 04:23:01 by Marti