

How to Perform a Wilcoxon Signed Rank Test and Analyze Paired Data

Authored by
stats writer

March 7, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Perform a Wilcoxon Signed Rank Test and Analyze Paired Data*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=134369>

Understanding the Wilcoxon Signed Rank Test

The **Wilcoxon Signed Rank Test** serves as a sophisticated statistical method designed to analyze **paired data** sets. In the realm of quantitative analysis, researchers often encounter scenarios where they must determine if a significant difference exists between two related groups. This test is particularly valuable because it evaluates the magnitude and direction of differences between pairs, rather than simply looking at the signs of those differences. By focusing on the **median** difference, it provides a clear picture of whether a specific intervention or condition has produced a measurable effect across a study population.

As a **non-parametric** statistical test, the Wilcoxon Signed Rank Test does not rely on the stringent assumption that the underlying data follows a **normal distribution**. This makes it an essential tool for scientists and analysts working with real-world data, which is frequently skewed or contains outliers that would violate the assumptions of parametric tests. Unlike the **paired t-test**, which requires the differences between pairs to be normally distributed, this test utilizes the ranking of differences to provide a **robust** alternative that maintains its validity even when the data is "messy."

The primary utility of this test lies in its ability to compare "before and after" measurements or "matched pairs." For example, a medical researcher might use the **Wilcoxon Signed Rank Test** to evaluate the effectiveness of a new medication by comparing the blood pressure of patients before and after treatment. Because it accounts for the relative size of the changes, it offers more **statistical power** than the simpler Sign Test, while remaining flexible enough to handle data that does not fit the classic "bell curve." Understanding how to correctly implement this test is vital for ensuring the integrity of **statistical inference** in various academic and professional fields.

When to Choose Non-Parametric Methods

Choosing the correct statistical test is a foundational step in any data analysis pipeline. You should opt for the **Wilcoxon Signed Rank Test** when you initially intended to use a **paired t-test**, but discovered that the distribution of the differences between your pairs is severely skewed. When the assumption of normality is violated, the results of a parametric test can become misleading, potentially leading to Type I or Type II errors. Therefore, the Wilcoxon test acts as a reliable safeguard, ensuring that your conclusions are based on the actual trends within the data rather than mathematical artifacts caused by distribution issues.

The most straightforward method to evaluate the distribution of your data is by generating a **histogram** of the calculated differences. If the resulting visual representation does not resemble a symmetric, bell-shaped curve, it is a strong indicator that a **non-parametric** approach is necessary. Researchers often look for heavy tails or significant gaps in the data that suggest the

normal distribution is an inappropriate model for the specific sample being studied. In such cases, the Wilcoxon test's reliance on ranks rather than raw values provides a necessary layer of protection against the influence of extreme data points.

However, it is important to remember that the **paired t-test** is relatively robust to minor departures from normality. This means that if your data is only slightly asymmetrical, the t-test might still be appropriate and offer slightly more power. The shift to a **Wilcoxon Signed Rank Test** is typically justified when the deviation from normality is severe or when the sample size is too small to confidently assess the distribution. By making this transition, you ensure that your **statistical significance** calculations remain accurate and defensible under peer review.

Core Assumptions and Data Requirements

Before proceeding with the calculation, one must ensure that the data meets specific criteria. First, the data must be paired, meaning each observation in one group corresponds to a specific observation in the second group. This is common in "repeated measures" designs where the same subjects are tested twice. Second, the differences between the pairs should be independent of one another. This **independence** ensures that the result of one pair does not influence the result of another, which is a fundamental requirement for the validity of the **p-value** calculation.

Another critical assumption for the **Wilcoxon Signed Rank Test** is that the differences between the pairs should come from a continuous distribution, and the distribution of these differences should be symmetric around the **median**. While the test does not require normality, this symmetry assumption is what allows us to test hypotheses regarding the median difference between the groups. If the distribution is highly asymmetric, the test can still be performed, but the interpretation shifts from a test of medians to a more general test of the **stochastic ordering** of the distributions.

The data itself should ideally be measured at the **interval or ratio level**, although the test can also be applied to **ordinal data** if the rankings are meaningful. Because the process involves calculating the magnitude of differences, the distance between values must be quantifiable. If your data consists of simple rankings without underlying numerical values, other tests might be more appropriate. Ensuring these data requirements are met at the outset will prevent errors during the interpretation phase of your study.

The Step-by-Step Mathematical Procedure

Performing the **Wilcoxon Signed Rank Test** involves a systematic seven-step process. The initial phase is data collection, where you gather pairs of related variables. For instance, in a clinical trial, these could be "pre-treatment" and "post-treatment" scores. Once the data is organized, you calculate the difference for each pair by subtracting the second value from the first. This produces a list of differences that can be positive, negative, or zero, representing the direction and scale of

change for each subject in the study.

The next phase involves ranking these differences based on their **absolute values**. During this step, you temporarily ignore the sign of the difference and focus solely on the magnitude. The smallest absolute difference receives a rank of 1, the second smallest a rank of 2, and so on. If any pairs have a difference of zero, they are typically excluded from the analysis, and the sample size is adjusted accordingly. After ranking, you reattach the original signs to these ranks, creating a set of **signed ranks**.

Finally, you calculate the test statistic, denoted as W , which is derived from the sum of the positive or negative ranks. This **test statistic** is then compared against a **critical value** obtained from a statistical table or software. By determining the **p-value**, you can assess the probability that the observed differences occurred by chance. If this value falls below your chosen **alpha level** (commonly 0.05), you can conclude that the difference between the groups is statistically significant.

Illustrative Case Study: A Basketball Training Scenario

To better understand the practical application of this test, consider a scenario involving a basketball coach who implements a new training program designed to improve free-throw accuracy. To evaluate the program's effectiveness, the coach selects 15 players and records the number of free throws each player successfully makes out of 20 attempts. These measurements are taken once before the training begins and once after the program is completed. This "pre-and-post" design creates perfectly **paired data**, as each data point in the "after" group is directly linked to a specific player in the "before" group.

Initially, the coach intended to use a **paired t-test** to analyze the results. However, upon reviewing the data, the coach observes that the differences in scores are not normally distributed. Perhaps a few players showed massive improvements while others showed none, creating a skewed distribution that makes the t-test unreliable. Consequently, the coach pivots to the **Wilcoxon Signed Rank Test** to obtain a more accurate assessment of whether the training program truly shifted the **median** performance of the team.

The following table presents the raw data collected from the 15 players. It serves as the foundation for our subsequent calculations, showing the performance of each individual before and after the intervention:

Player	Before	After
Player #1	14	15
Player #2	17	17
Player #3	12	15
Player #4	15	15
Player #5	15	17
Player #6	9	14
Player #7	12	9
Player #8	13	14
Player #9	13	11
Player #10	15	16
Player #11	19	18
Player #12	17	20
Player #13	14	20
Player #14	14	10
Player #15	16	17

Defining the Null and Alternative Hypotheses

Every statistical analysis must begin with a clear definition of the hypotheses being tested. In the context of the basketball example, the **null hypothesis** (H_0) states that there is no significant difference between the two groups. Specifically, it assumes that the **median** difference between the "before" and "after" scores is zero. Essentially, this hypothesis suggests that any observed improvements in free-throw shooting are merely the result of random chance or **sampling error** rather than the training program itself.

Conversely, the **alternative hypothesis** (H_A) represents the claim the coach is trying to prove. In this specific case, the coach is looking for a directional improvement, so the hypothesis states that the median difference is negative (meaning players scored significantly lower before the training than they did after). This is known as a one-tailed test. Defining these hypotheses clearly is essential because it dictates how we will eventually interpret the **p-value** and whether we will reject the null hypothesis in favor of the coach's training program.

It is important to select the **significance level**, or alpha, at this stage as well. Most social and athletic studies utilize an alpha of 0.05. This means the coach is willing to accept a 5% risk of concluding that the training program works when, in reality, it does not. By setting these parameters before looking at the final results, the researcher maintains **scientific objectivity** and prevents the temptation to adjust the criteria to fit the data.

Processing Data: Differences and Absolute Values

With the hypotheses established, the coach proceeds to the computational phase. The first task is to calculate the difference (d_i) for each pair of scores (Before - After). If a player made 12 shots before and 15 shots after, the difference is -3. If another player made 10 shots before and 10 after, the difference is 0. These raw differences tell us exactly how much each player's performance changed. Following this, the coach must calculate the **absolute difference** for each pair, which effectively strips the direction of the change and leaves only the magnitude.

The calculation of absolute differences is a crucial step in the **Wilcoxon Signed Rank Test** because it allows for the ranking of changes regardless of whether they were improvements or declines. This focuses the test on the "strength" of the change. In our basketball dataset, this transformation looks like this:

Player	Before	After	Difference	Abs. Difference
Player #1	14	15	-1	1
Player #2	17	17	0	0
Player #3	12	15	-3	3
Player #4	15	15	0	0
Player #5	15	17	-2	2
Player #6	9	14	-5	5
Player #7	12	9	3	3
Player #8	13	14	-1	1
Player #9	13	11	2	2
Player #10	15	16	-1	1
Player #11	19	18	1	1
Player #12	17	20	-3	3
Player #13	14	20	-6	6
Player #14	14	10	4	4
Player #15	16	17	-1	1

This systematic approach ensures that every player's change is accounted for. By visualizing these differences, the coach can begin to see whether the majority of players improved (indicated by negative differences) or if the results are scattered. However, the raw differences alone are not enough to determine **statistical significance**; they must be transformed into ranks to satisfy the non-parametric nature of the test.

The Ranking Protocol and Addressing Ties

The ranking process is where the **Wilcoxon Signed Rank Test** derives its name. All absolute

differences are ordered from smallest to largest. The smallest non-zero difference is assigned a rank of 1. If multiple pairs have the same absolute difference, they are considered "ties." In such cases, you must assign the **mean** of the ranks that would have been assigned to those observations. For example, if two pairs are tied for the 3rd and 4th positions, both are assigned a rank of 3.5. This ensures that the total sum of ranks remains consistent.

During this stage, any pairs with an absolute difference of zero are entirely excluded from the ranking process. These cases indicate no change, and including them would skew the **test statistic**. In our basketball example, there were 15 players, but if two players had no change in their scores, the effective sample size (n) for the test would drop to 13. This adjustment is vital for accurately looking up the **critical value** later in the process.

Once the ranks are assigned, the original signs (+ or -) are restored to the ranks. If a difference was negative, its corresponding rank becomes negative. This results in a list of "signed ranks" that reflect both the magnitude of the change and its direction. The following images demonstrate how these ranks are assigned and how the final signed ranks are organized for the basketball dataset:

Player	Before	After	Difference	Abs. Difference	Rank
Player #2	17	17	0	0	-
Player #4	15	15	0	0	-
Player #1	14	15	-1	1	3
Player #8	13	14	-1	1	3
Player #10	15	16	-1	1	3
Player #11	19	18	1	1	3
Player #15	16	17	-1	1	3
Player #5	15	17	-2	2	6.5
Player #9	13	11	2	2	6.5
Player #3	12	15	-3	3	9
Player #7	12	9	3	3	9
Player #12	17	20	-3	3	9
Player #14	14	10	4	4	11
Player #6	9	14	-5	5	12
Player #13	14	20	-6	6	13

Player	Before	After	Difference	Abs. Difference	Rank	Negative Ranks	Positive Ranks
Player #2	17	17	0	0	-		
Player #4	15	15	0	0	-		
Player #1	14	15	-1	1	3	-3	
Player #8	13	14	-1	1	3	-3	
Player #10	15	16	-1	1	3	-3	
Player #11	19	18	1	1	3		3
Player #15	16	17	-1	1	3	-3	
Player #5	15	17	-2	2	6.5	-6.5	
Player #9	13	11	2	2	6.5		6.5
Player #3	12	15	-3	3	9	-9	
Player #7	12	9	3	3	9		9
Player #12	17	20	-3	3	9	-9	
Player #14	14	10	4	4	11		11
Player #6	9	14	-5	5	12	-12	
Player #13	14	20	-6	6	13	-13	
Sum						-61.5	29.5

Deriving the Test Statistic (W)

The final numerical objective of the **Wilcoxon Signed Rank Test** is to calculate the test statistic, typically represented by the letter **W**. This value is determined by summing the positive ranks and the negative ranks separately. The test statistic **W** is defined as the smaller of the two absolute sums. By identifying the smaller sum, the test provides a single value that characterizes the degree of overlap between the improvements and the declines within the sample.

In our basketball case study, after summing the ranks, we find that the smaller of the two sums is 29.5. This means our calculated test statistic is **W = 29.5**. This value represents the cumulative weight of the differences that oppose the general trend. If the training program were incredibly effective, we would expect a very small **W** value, as almost all ranks would be in one direction (the direction of improvement), leaving the "opposing" sum near zero.

Understanding the **W** statistic is key to **statistical inference**. Because the distribution of **W** under the null hypothesis is known, we can determine how likely it is to see a **W** as small as 29.5 by chance. Unlike many other tests where a "larger" statistic indicates significance, in the Wilcoxon test, a smaller **W** relative to the sample size suggests a more significant result. This reflects a clear dominance of one sign over the other in the ranked data.

Statistical Significance and Critical Value Assessment

To determine whether to reject the **null hypothesis**, the coach must compare the calculated **W** statistic to a **critical value**. This critical value is found in a standard Wilcoxon Signed Rank Table,

using the adjusted sample size ($n = 13$) and the chosen **alpha level** of 0.05. The table provides a threshold; if the calculated W is less than or equal to this threshold, the results are considered statistically significant.

For this specific dataset, the critical value for $n = 13$ at an alpha of 0.05 is **17**. Since the coach's calculated W (29.5) is **not** less than or equal to 17, the result fails to reach the threshold for **statistical significance**. The visual reference for this critical value determination is shown below:

	Alpha value				
n	0.005	0.01	0.025	0.05	0.10
5	-	-	-	-	0
6	-	-	-	0	2
7	-	-	0	2	3
8	-	0	2	3	5
9	0	1	3	5	8
10	1	3	5	8	10
11	3	5	8	10	13
12	5	7	10	13	17
13	7	9	13	17	21
14	9	12	17	21	25
15	12	15	20	25	30
16	15	19	25	29	35
17	19	23	29	34	41
18	23	27	34	40	47
19	27	32	39	46	53
20	32	37	45	52	60
21	37	42	51	58	67
22	42	48	57	65	75
23	48	54	64	73	83
24	54	61	72	81	91
25	60	68	79	89	100
26	67	75	87	98	110
27	74	83	96	107	119
28	82	91	105	116	130
29	90	100	114	126	140
30	98	109	124	137	151

As a result, the coach must conclude that there is insufficient evidence to suggest that the training program led to a significant increase in free-throw performance. While some players may have improved, the overall shift across the team was not strong enough to rule out random variation. This outcome demonstrates the **Wilcoxon Signed Rank Test** as a conservative and reliable judge of data, preventing researchers from making overconfident claims based on weak or inconsistent evidence.

Summary and Final Interpretations

In summary, the **Wilcoxon Signed Rank Test** is an indispensable tool for analyzing **paired data** when the assumptions of parametric tests cannot be met. By utilizing the **ranking** of differences rather than their raw values, it provides a robust method for testing whether a significant change has occurred between two related groups. Its ability to handle non-normal distributions and mitigate the influence of outliers makes it a preferred choice in clinical trials, psychological studies, and athletic performance analysis.

The process of performing the test--from stating the **null hypothesis** to comparing the test statistic against a **critical value**--requires careful attention to detail, particularly when dealing with ties and zero differences. However, the clarity it provides is well worth the effort. In the basketball example, it helped the coach make a data-driven decision, avoiding the mistake of assuming a program was effective just because a few players showed raw improvements.

Whether you are a student, a researcher, or a professional analyst, mastering the **Wilcoxon Signed Rank Test** enriches your statistical toolkit. It ensures that your findings are grounded in **robust mathematical principles** and that your interpretations of paired data are both accurate and scientifically sound. For those looking to streamline this process, modern software or a dedicated calculator can perform these steps automatically, allowing you to focus on the broader implications of your research.

Note: Use the **Wilcoxon Signed Rank Test** if you wish to perform the test using a calculator instead of by hand.