Is there a difference between two paired samples?

Authored by

stats writer

December 26, 2025

When conducting statistical analysis, distinguishing between different types of data structures is critical for selecting the appropriate test. The concept of **paired samples** arises when observations in one group are naturally linked, or dependent, on observations in the second group. This linkage is often due to measuring the same subject twice—such as "before" and "after" an intervention—or matching two subjects based on specific characteristics.

The fundamental statistical question concerning **paired samples** is whether a systematic change or difference exists between the two measurements. This difference is not merely the difference between the aggregate means, but rather the mean of the individual differences calculated for each pair. If the intervention had no effect, we would expect the mean difference to be close to zero. The goal of the analysis, therefore, is to quantify this difference and determine its **statistical significance**.

Determining the significance requires utilizing a test designed specifically for dependent data. If we assume that the distribution of these differences is approximately **normal**, the standard choice is the **paired t-test**. However, real-world data frequently violates this assumption, particularly when sample sizes are small or when the underlying phenomenon naturally produces skewed results. When the assumption of normality is untenable, statisticians must turn to alternative methods that do not rely on strict distributional parameters, leading us directly to the realm of **non-parametric** statistics.

When Parametric Tests Fail: Introducing the Wilcoxon Signed-Rank Test

The standard **paired t-test** is a powerful tool, but its validity rests heavily on the premise that the distribution of the differences between the paired observations follows a **normal distribution**. When this assumption is violated, the reliability of the t-test's results, particularly the calculated **p-value**, can be severely compromised, potentially leading to incorrect conclusions regarding the true effect of the intervention. This is a common challenge, especially in clinical trials or focused behavioral studies where sample populations are often limited.

In response to this limitation, the **Wilcoxon Signed-Rank Test** was developed. This test serves as the leading **non-parametric** counterpart to the **paired t-test**. Crucially, it does not require the distribution of the differences to be normal. Instead of analyzing the magnitude of the differences directly, it focuses on the ranks and signs of those differences. This reliance on ranks makes the test robust against outliers and violations of distributional assumptions.

The core mechanism of the **Wilcoxon Signed-Rank Test** involves calculating the difference score for each pair, ranking the absolute values of these differences, and then summing the ranks separately for positive and negative differences. The test statistic, usually denoted as W or V, is derived from these sums. By analyzing these ranked scores, we can determine if the positive differences tend to be larger (in rank) than the negative differences, suggesting a true shift in the

population's central tendency. This methodology allows researchers to draw reliable inferences about location shifts even when facing data that is highly skewed or non-normally distributed.

Key Assumptions and Hypotheses of the Test

While the **Wilcoxon Signed-Rank Test** frees us from the constraints of the normality assumption, it is not entirely assumption-free. The primary requirement is that the data must come from a continuous distribution, and the paired differences must be symmetrically distributed around the median. If the distribution of differences is highly asymmetric, an alternative test, such as the Sign Test, might be more appropriate, although the Wilcoxon test is generally considered more powerful. Additionally, the data must be measured on at least an ordinal scale, allowing for meaningful ranking of the differences.

The hypotheses underlying the **Wilcoxon Signed-Rank Test** are framed around the median difference (eta_D). Unlike the t-test which focuses on the mean, the Wilcoxon test focuses on the median because of its resistance to extreme values in non-normal distributions. The test aims to determine if the median difference between the paired observations is zero.

For a two-sided test, the hypotheses are formally defined as:

Null Hypothesis ($text{H}_0$): The median difference between the paired observations is zero ($eta_D = 0$). This suggests that the training program or intervention had no measurable effect.

Alternative Hypothesis ($text{H}_A$): The median difference between the paired observations is not zero ($eta_D neq 0$). This suggests a statistically significant difference exists, but the direction (increase or decrease) is not specified.

The selection of the appropriate alternative hypothesis--two-sided, left-tailed ("less"), or right-tailed ("greater")--is critical and should be dictated by the research question established before data collection. This careful formulation ensures that the statistical analysis directly addresses the researcher's intended inquiry regarding the presence and potential direction of the effect.

Case Study: Evaluating a Basketball Training Program

To illustrate the practical application of the **Wilcoxon Signed-Rank Test**, consider a scenario involving athletic performance improvement. A dedicated basketball coach wants to rigorously evaluate whether a newly implemented, intensive four-week training regimen successfully increases the free-throw accuracy of his players. To isolate the effect of the program, he recruits 15 players for the study.

The experimental design relies on **paired samples**: each of the 15 players acts as their own control. Before the training program begins, each player attempts 20 free throws, and their

successful shots are recorded. Following the completion of the four-week program, the same 15 players repeat the test, again attempting 20 free throws. The resulting data structure is intrinsically dependent, as the "after" score is directly linked to the "before" score for Player 1, Player 2, and so on.

Initially, the coach intended to use a **paired t-test**, the most common statistical tool for this type of dependent data. However, upon analyzing the distribution of the difference scores (After - Before) across the 15 players, he observed significant skewness and heavy tails, indicating a clear deviation from the required **normal distribution** assumption. Recognizing this violation, the coach correctly shifted his methodology to the more robust **non-parametric** alternative: the **Wilcoxon Signed-Rank Test**.

Data Visualization and Preparation

The integrity of any statistical conclusion relies on having transparent and well-organized data. In this basketball example, the raw data consists of two columns: the number of free throws made before training and the number made after training. This structure allows us to immediately calculate the key variable for the Wilcoxon test: the difference score. It is often helpful to visualize this raw data to intuitively understand the magnitude and direction of the changes, although the formal test relies on ranks.

The following table summarizes the performance data (out of 20 attempts) for the 15 players, clearly demonstrating the paired nature of the observations. This dataset must be loaded into the statistical software, in this case, **R**, as two distinct vectors or columns, ensuring that the observation order remains consistent across the vectors to preserve the pairing.

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

Before running the test, data preparation in **R** involves defining these two sets of scores as separate, equal-length vectors. If the pairing were inadvertently broken (e.g., if the scores were shuffled), the resulting Wilcoxon test would be invalid. The statistical calculation relies entirely on the subtraction of the corresponding paired values before ranking the absolute differences. This meticulous preparation is foundational to achieving a valid statistical inference.

Conducting the Wilcoxon Signed-Rank Test in R

The statistical programming language **R** provides a highly efficient and straightforward method for executing the **Wilcoxon Signed-Rank Test** using the built-in function `wilcox.test()`. For this function to correctly perform the paired analysis, specific arguments must be passed, ensuring that the software recognizes the dependency between the two input vectors.

The general syntax for the paired version of the test is:

wilcox.test(x, y, paired=TRUE)

In this syntax:

x, y: These represent the two vectors containing the data values. It is customary to input the 'Before' scores as the first vector (x) and the 'After' scores as the second vector (y) to maintain consistent interpretation of the signs of the differences (y - x).

paired: Setting this argument to **TRUE** is absolutely essential. This command instructs **R** to calculate the differences internally and apply the Wilcoxon Signed-Rank ranking methodology, rather than treating the samples as independent.

The following code illustrates how to use this function to perform the Wilcoxon Signed-Rank Test on this data:

#create the two vectors of data
before <- c(14, 17, 12, 15, 15, 9, 12, 13, 13, 15, 19, 17, 14, 14, 16)
after <- c(15, 17, 15, 15, 17, 14, 9, 14, 11, 16, 18, 20, 20, 10, 17)

#perform Wilcoxon Signed-Rank Test
wilcox.test(before, after, paired=TRUE)

Wilcoxon signed rank test with continuity correction

data: before and after
V = 29.5, p-value = 0.275
alternative hypothesis: true location shift is not equal to 0

Interpreting the Two-Sided Test Results

The output generated by **R** provides the necessary components to draw a formal conclusion about the efficacy of the basketball training program. The critical components are the test statistic V and the calculated **p-value**. In this specific execution, the output indicates that the test statistic V is **29.5**, and the corresponding **p-value** is **0.275**.

The decision-making process in hypothesis testing involves comparing the **p-value** to a pre-determined level of significance, denoted as $alpha$ (alpha). Conventionally, researchers set $alpha$ at 0.05, representing a 5% risk of incorrectly rejecting a true **null hypothesis** (Type I error). The rule is straightforward: if $p le alpha$, we reject the null hypothesis; if $p > alpha$, we fail to reject the null hypothesis.

In our case, the calculated **p-value** (0.275) is substantially greater than the conventional significance level of 0.05. Therefore, we **fail to reject the null hypothesis**. The conclusion must be phrased cautiously: based on the evidence provided by this sample and analyzed using the **Wilcoxon Signed-Rank Test**, there is insufficient statistical evidence to conclude that the training program resulted in a statistically significant shift in the players' free-throw performance. The observed differences could reasonably be attributed to random variation or chance.

Utilizing One-Sided (Directional) Hypotheses

While the two-sided test addresses whether a difference exists in either direction, researchers often have a strong theoretical basis or prior evidence suggesting the direction of the expected change. For instance, the basketball coach likely hypothesized that the training program would specifically *increase* the number of free throws made. In such scenarios, using a one-sided (or directional) alternative hypothesis provides a more focused and potentially powerful test.

The `wilcox.test()` function in **R** allows for the specification of directional tests using the `alternative` argument. If we hypothesize that the "after" scores are significantly *less* than the "before" scores (i.e., the training harmed performance), we use `alternative="less"`. Conversely, if we hypothesize that the "after" scores are significantly *greater* than the "before" scores (i.e., the training improved performance), we use `alternative="greater"`. Note that "less" corresponds to a negative median shift, and "greater" corresponds to a positive median shift, usually defined as $y - x$.

Applying these directional tests to the basketball data provides a deeper view. The output below shows the results for both potential one-sided scenarios. The resulting **p-value** for the left-tailed test (0.1375) indicates that we cannot conclude performance decreased. More importantly, the right-tailed test, reflecting the coach's expectation of improvement, yields a very high **p-value** (0.8774), confirming that there is no evidence to support the claim that the median difference is greater than zero, even under a directional hypothesis structure.

#perform left-tailed Wilcoxon Signed-Rank Test
wilcox.test(before, after, paired=TRUE, alternative="less")

Wilcoxon signed rank test with continuity correction

data: before and after
V = 29.5, p-value = 0.1375
alternative hypothesis: true location shift is less than 0

#perform right-tailed Wilcoxon Signed-Rank Test
wilcox.test(before, after, paired=TRUE, alternative="greater")

Wilcoxon signed rank test with continuity correction

data: before and after
V = 29.5, p-value = 0.8774
alternative hypothesis: true location shift is greater than 0

Conclusion: Practical Applications of Non-Parametric Testing

The analysis of the basketball training data underscores the critical importance of selecting the appropriate statistical tool. By correctly identifying that the difference scores violated the **normal distribution** assumption, the coach avoided the potential pitfalls of the **paired t-test** and instead utilized the robust **Wilcoxon Signed-Rank Test**. This methodological rigor ensures that the conclusion—that the program did not cause a statistically significant change—is reliable, regardless of the non-normal shape of the underlying data distribution.

The application of the **Wilcoxon Signed-Rank Test** is widespread across disciplines where small sample sizes and non-normal data are common, including psychology, biology, and experimental sciences. It is the definitive choice for analyzing dependent data when distributional assumptions cannot be met, providing a powerful means to test for location shifts using ranks rather than raw scores. This reliance on ranking minimizes the influence of extreme scores, offering a safer path to inference when data conditions are challenging.

In summary, understanding the nuances of **paired samples** is essential. Yes, there is a difference between two paired samples, measured by the individual differences between pairs. The challenge lies in determining if this calculated difference is statistically meaningful, which is expertly handled by the **Wilcoxon Signed-Rank Test** when parametric assumptions fail. Mastering the implementation of this test in software like **R** allows practitioners to conduct rigorous, assumption-light statistical analyses on dependent data, ensuring sound conclusions in empirical research.