How to conduct a Wilcoxon Signed-Rank Test in Python?

Authored by

stats writer

December 25, 2025

The Wilcoxon Signed-Rank Test is a cornerstone statistical tool, classified as a non-parametric test. Its primary utility lies in comparing two related samples or repeated measurements on a single sample, aiming to determine if their population distributions differ significantly. Unlike its parametric counterpart, the paired t-test, the Wilcoxon test does not require the assumption that the differences between paired observations are normally distributed, making it highly robust for skewed or ordinal data.

Implementing this powerful statistical analysis in Python is streamlined through the SciPy library, specifically using the `scipy.stats.wilcoxon()` function. This function efficiently processes the two input samples and calculates the crucial output metrics: the W statistic (or T statistic, depending on the variant calculation) and the corresponding p-value. These metrics are fundamental for hypothesis testing, allowing researchers to formally assess whether the observed differences between the two related groups are statistically significant or merely due to random chance.

Introduction to the Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test serves as the essential non-parametric alternative to the standard paired t-test. It is specifically designed for situations involving **dependent samples**--where measurements are taken from the same subjects under two different conditions or from matched pairs. The test evaluates the null hypothesis that the median difference between the paired observations is zero, implying that the underlying population distributions are identical, or at least symmetric around zero.

This test is particularly valued in disciplines like psychology, medicine, and engineering when assumptions regarding data distribution cannot be met. For instance, if data consists of scores or ratings (ordinal data), or if the sample size is small, resulting in a questionable assumption of normality, the Wilcoxon test provides a statistically sound method to proceed. By focusing on the ranks of the differences rather than the differences themselves, the test mitigates the influence of potential outliers, providing a more robust measure of location shift.

In essence, the Wilcoxon Signed-Rank Test is used precisely to test whether or not there is a significant difference in the location parameter (usually the median) of the populations from which the two dependent samples were drawn, especially when the normality assumption--critical for the paired t-test--is violated or cannot be reasonably assumed. The ability to handle non-normal distributions without sacrificing too much statistical power makes it an indispensable tool in the statistical toolbox.

Why Choose a Non-Parametric Test?

The choice between parametric and non-parametric statistics often hinges on the characteristics of

the data and the validity of underlying assumptions. Parametric tests, such as the paired t-test, rely heavily on the assumption that the data (or, more precisely, the differences between pairs) follow a **normal distribution**. When this condition is satisfied, parametric tests typically offer greater statistical power, meaning they are more likely to correctly reject a false null hypothesis.

However, real-world data frequently deviates from the ideal normal distribution. Data might be heavily skewed, contain substantial outliers, or represent ordinal scale measurements where the distance between points is not consistent. In such scenarios, applying a parametric test can lead to inaccurate p-value calculations and unreliable conclusions. This is where non-parametric tests, like the Wilcoxon Signed-Rank Test, become essential.

A key advantage of the Wilcoxon test is that it operates on the **ranks of the data differences**, rather than the raw numerical values. By ranking the absolute differences and then summing the ranks based on the sign (positive or negative) of the original difference, the test effectively removes the dependence on specific distribution shapes. This procedure ensures that the statistical inference remains valid even when the underlying data distributions are complex or unknown, thus safeguarding the integrity of the research findings.

Understanding the Hypotheses and Assumptions

Before executing the Wilcoxon Signed-Rank Test, it is crucial to clearly define the formal statistical hypotheses being tested. Since this test addresses paired observations, the hypotheses focus on the median of the difference scores between the two groups. The test fundamentally assesses whether the distribution of differences is centered at zero.

The standard hypotheses for a two-sided test are defined as follows:

H0: The Null Hypothesis: The distribution of the differences between the paired observations is symmetric about zero (i.e., the median difference is zero). This implies that there is **no systematic difference** between the two conditions or measurements.
HA: The Alternative Hypothesis: The distribution of the differences is **not** symmetric about zero (i.e., the median difference is not zero). This suggests that the treatment or intervention has caused a statistically significant shift in the measured outcome.

While the Wilcoxon Signed-Rank Test is highly flexible as a non-parametric test, it is not entirely assumption-free. The core assumptions required for valid inference include: 1) The data must come from **dependent (paired) observations**. 2) The variable under investigation must be measured on at least an ordinal scale, allowing for meaningful ranking. 3) The distribution of the differences must be symmetric around the median under the null hypothesis, although the strict normality assumption is dropped.

The Mechanics of the Test Statistic (W)

The calculation of the Wilcoxon Signed-Rank Test statistic, often denoted as **W** or **T**, is based on a structured process involving the differences between the paired samples. This process provides the foundation for determining the probability associated with the observed data, assuming the null hypothesis is true.

The calculation involves several key steps:

Calculate Differences: The difference (D_i) between each pair of observations is calculated ($y_i - x_i$).
Ignore Zero Differences: Pairs where the difference is exactly zero are typically removed from the analysis, and the sample size (n) is reduced accordingly.
Determine Absolute Differences and Ranks: The absolute values of these differences ($\vert D_i \vert$) are then calculated and ranked from smallest to largest. Ties in absolute differences receive the average of the ranks they span.
Apply Signs to Ranks: Finally, the sign (positive or negative) of the original difference (D_i) is applied back to the corresponding rank.

The test statistic **W** is then derived by summing the ranks associated with the positive differences (or the minimum of the sum of positive ranks and the sum of negative ranks, depending on the chosen definition). A small value of W (or a value far from the expected mean under H0) indicates a significant deviation from the expectation that positive and negative differences should balance out. This calculated W statistic is then used to determine the associated p-value, which dictates the decision regarding the null hypothesis.

Practical Example: Fuel Treatment Efficacy Study

To illustrate the application of this test, consider a common scenario in automotive engineering research. Researchers are interested in evaluating whether a newly developed fuel treatment significantly alters the average miles per gallon (mpg) performance of a specific type of car. This study design necessitates a **paired comparison** because the same set of vehicles must be tested both before and after the treatment is applied, controlling for vehicle variability.

In this hypothetical study, 12 identical cars are selected. The mpg is first measured under standard conditions (Group 1). Then, the new fuel treatment is applied, and the mpg is measured again for the same 12 cars (Group 2). The goal is to perform the Wilcoxon Signed-Rank Test in Python to ascertain if there is a statistically significant difference in the mean mpg between the untreated and treated conditions.

We use the following steps to perform the analysis. Since we cannot guarantee that the distribution

of differences in mpg follows a normal distribution, the non-parametric approach is the most appropriate and rigorous methodology to determine if there is a difference in the mean mpg between the two groups.

Step 1: Setting Up the Data in Python

The foundational element of any statistical analysis is accurately structuring the input data. For the Wilcoxon Signed-Rank Test, the data must be organized into two corresponding arrays or lists, where the elements at the same index represent the paired observations from the same unit. We begin by creating two arrays to house the mpg values collected for each group of the 12 sampled vehicles:

group1 =
group2 =

It is critical to ensure that the order of observations in **`group1`** corresponds directly to the order of observations in **`group2`**. For example, the first element (20) in `group1` represents the mpg of Car 1 before treatment, and the first element (24) in `group2` represents the mpg of Car 1 after treatment. The two arrays must be of equal length, satisfying the requirement for dependent samples testing.

Step 2: Executing the Wilcoxon Test using SciPy

Once the data is correctly structured, the next phase involves importing the necessary libraries and executing the statistical test. The SciPy library provides the robust **`stats.wilcoxon()`** function specifically tailored for this analysis. We must import the `scipy.stats` module to access this functionality.

The general syntax for the Wilcoxon function is highly intuitive and accepts the following primary parameters:

```
wilcoxon(x, y, alternative='two-sided')
```

Where the arguments are interpreted as:

x: An array of sample observations from group 1 (e.g., the baseline measurements).
y: An array of sample observations from group 2 (e.g., the post-treatment measurements).
alternative: Defines the alternative hypothesis. The default is `'two-sided'`, testing for any difference, but other options include `'less'` and `'greater'` for one-sided tests.

Here is how to use this function in our specific example, followed by the resulting output:

import scipy.stats as stats

```
#perform the Wilcoxon-Signed Rank Test
stats.wilcoxon(group1, group2)

(statistic=10.5, pvalue=0.044)
```

The execution yields a test statistic (W) of **10.5** and a corresponding two-sided p-value of **0.044**.

Interpreting the Results and Drawing Conclusions

The final and most critical step in hypothesis testing is interpreting the calculated p-value in the context of the defined null hypothesis. For this example, we established the following framework:

H0: The median mpg is equal between the two groups (i.e., the fuel treatment has no effect).

HA: The median mpg is **not** equal between the two groups (i.e., there is a significant effect).

The calculated p-value of **0.044** represents the probability of observing our current data (or data more extreme) if the null hypothesis were actually true. We compare this against a conventional significance level ($alpha$) of 0.05.

Since the p-value (**0.044**) is less than the significance level (0.05), we **reject the null hypothesis**. This rejection implies that we have sufficient statistical evidence to conclude that the fuel treatment did cause a statistically significant change in the true median mpg performance of the cars. The difference observed is unlikely to be due merely to random variation.