Learn How to Perform a Wilcoxon Signed-Rank Test in Python

Authored by

Mohammed looti

November 8, 2025

The Wilcoxon Signed-Rank Test stands out as an exceptionally powerful tool within non-parametric statistics, specifically designed for analyzing data derived from **dependent** or **paired samples**. It provides a robust, statistically sound alternative to the traditional paired t-test, particularly when the stringent requirements of parametric testing--most notably the assumption of normality in difference scores--cannot be reliably satisfied or verified. This test is indispensable in fields ranging from clinical research to engineering, wherever comparative measurements are taken on the same subjects or matched pairs.

The primary function of this specialized procedure is to rigorously assess whether a statistically significant discrepancy exists between the distributions of two related population groups. Unlike parametric methods that rely heavily on population means, the Wilcoxon test operates by analyzing the **ranks** of the observed differences, rather than the raw scores themselves. This methodology grants the test significant resilience against common data issues such as extreme outliers and severely skewed distributions, ensuring the derived inference remains reliable even when data assumptions are violated.

This comprehensive guide is structured to demystify the core mathematical principles underlying the Wilcoxon Signed-Rank Test. Furthermore, it delivers a meticulous, step-by-step tutorial demonstrating how to execute this critical analysis effectively using the statistical capabilities provided by Python's leading scientific library, scipy.stats. Mastering this technique is essential for any data analyst or researcher dealing with within-subject experimental designs.

Fundamentals of the Wilcoxon Signed-Rank Test

The application of the Wilcoxon Signed-Rank Test is strictly limited to scenarios involving **dependent samples**. This crucial prerequisite means that the observations collected in the first measurement condition must be inherently linked or correlated with the observations collected in the second condition. Classic research designs that necessitate this test include longitudinal studies where measurements are taken "before and after" an intervention on the identical subjects, or highly controlled experiments utilizing tightly matched pairs based on specific demographic or physical characteristics. This internal structure of dependency is the defining feature that differentiates the Wilcoxon Signed-Rank Test from its counterpart for independent groups, the Mann-Whitney U Test.

The operational mechanics of the test are ingenious and robust. The process begins by calculating the precise **magnitude and direction** (positive or negative sign) of the difference observed between each pair of dependent observations. Subsequently, the test procedure ranks the absolute values of these calculated differences, ignoring the sign initially. The final and pivotal step involves summing the ranks that correspond specifically to the positive differences (R+) and summing the ranks associated with the negative differences (R-). The resulting test statistic,

typically denoted as W, is derived from comparing these two summed ranks. If the positive and negative differences are relatively balanced, it suggests the intervention had no systematic effect; conversely, a significant skew in the rank sums indicates a substantial treatment effect.

When applying this test within the framework of statistical hypothesis testing, the inferential focus shifts distinctly away from comparing population means--the primary goal of the t-test--towards comparing population **medians**. The primary question addressed by the Wilcoxon test is whether the population median of the differences between the paired observations is significantly dissimilar from zero. By focusing on the median, which is less sensitive to extreme values, the test remains highly reliable even when researchers cannot confidently assume either an interval scale for the data or the prerequisite assumption of normality. This makes it a preferred non-parametric choice for ensuring valid statistical inference in challenging data environments.

Designing the Paired Study: A Practical Example

Consider a practical scenario within the automotive research sector. Engineers are tasked with determining if a newly developed chemical additive for fuel can induce a statistically significant alteration in the average miles per gallon (mpg) achieved by a specific fleet vehicle model. To achieve maximum control and isolate the additive's effect, they implement a classic **within-subjects design**: they measure the mpg performance of a small, distinct set of 12 vehicles both before the additive is introduced (baseline condition) and again after the additive has been applied (experimental condition).

The inherent experimental structure dictates that the data collected is fundamentally **dependent**, as the "before" and "after" measurements are taken on the identical 12 cars. This pairing is crucial for the analysis. Furthermore, due to the limited sample size (N=12), establishing or maintaining the assumption that the scores for the differences between the paired readings follow a normal distribution becomes statistically tenuous. Consequently, relying on a parametric test, such as the paired t-test, would introduce unacceptable risk of error.

Given these data characteristics and design constraints, the Wilcoxon Signed-Rank Test emerges as the most appropriate and statistically sound methodological choice. This tutorial will proceed by utilizing a structured analytical procedure in Python. The objective is to determine definitively, based on rigorous statistical evidence, whether the fuel treatment introduced a statistically relevant difference in the median mpg achieved when comparing the baseline condition against the post-treatment experimental condition.

Python Implementation: Data Preparation and Execution

Step 1: Defining and Structuring the Dependent Dataset. The initial, crucial step in any statistical analysis using Python involves accurately defining the dataset. We must create two

distinct arrays or lists to hold the paired mpg values for the baseline and post-treatment groups. It is absolutely essential that the positional order of observations is perfectly preserved across both lists; this ensures that the difference calculation correctly pairs the "before" score of Car 1 with the "after" score of Car 1, and so on:

group1 =
group2 =

Step 2: Executing the Wilcoxon Signed-Rank Test using SciPy. With the data correctly structured, we can proceed directly to the statistical computation. We utilize the powerful `wilcoxon` function, which is readily available within the scipy.stats library. This function is engineered to efficiently manage all internal calculations, including the complex ranking and summing procedures required by the test. The generalized structure for invoking the function is defined clearly below:

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

The necessary input parameters required for successful execution within the Python environment are detailed as follows:

x: This array contains the sample observations collected under the first condition, representing the baseline measurements (e.g., mpg without the fuel treatment).
y: This array holds the sample observations corresponding to the second, dependent condition, representing the post-treatment measurements (e.g., mpg after applying the fuel additive).
alternative: This critical, optional parameter dictates the specific focus of the alternative hypothesis (HA). The default configuration is `'two-sided'`, which tests for any non-zero difference between the distributions. Researchers can specify `'less'` if their hypothesis predicts that X is stochastically greater than Y, or `'greater'` if they hypothesize that X is stochastically less than Y.

The specific Python code implementation required to run the Wilcoxon Signed-Rank Test for our automotive fuel efficiency example is presented here, followed by the resulting output tuple:

import scipy.stats as stats

# Perform the Wilcoxon Signed-Rank Test on the paired data arrays
stats.wilcoxon(group1, group2)

(statistic=10.5, pvalue=0.044)

The execution successfully returns two key metrics: a Wilcoxon test statistic (W) of **10.5**, and a corresponding two-sided p-value of **0.044**. These numerical results form the indispensable foundation for reaching a definitive statistical conclusion regarding the efficacy of the fuel treatment

additive.

Analyzing Results and Drawing Statistical Conclusions

To correctly interpret the statistical output generated by SciPy, we must first formally define the two competing statements central to our hypothesis testing framework: the null hypothesis (H0) and the alternative hypothesis (HA). For this specific paired comparison concerning fuel efficiency:

H0 (Null Hypothesis): The median difference in mpg between the two groups (before and after treatment) is exactly zero. This posits that the fuel treatment additive yields no measurable median effect on vehicle efficiency.
HA (Alternative Hypothesis): The median difference in mpg between the two groups is *not* zero. This asserts that the fuel treatment results in a statistically detectable change in median efficiency.

The definitive step in drawing a statistical conclusion involves comparing the calculated p-value against the predetermined significance level, denoted as $alpha$ ($alpha$). The industry standard significance threshold is almost universally set at $alpha = 0.05$. The decision rule is straightforward: if the calculated p-value is less than or equal to the chosen $alpha$ level, we must **reject the null hypothesis**, thereby concluding that the observed effect is statistically significant and unlikely to be due to random chance alone.

In our example, the resulting p-value is **0.044**. Since 0.044 is strictly less than the standard significance level of 0.05, we have sufficient empirical evidence to firmly reject the null hypothesis (H0). Based on the non-parametric analysis, we can confidently conclude that the true population median mpg differs significantly between the baseline condition and the post-treatment condition. In practical terms relevant to the automotive engineers, the new fuel treatment additive has successfully introduced a statistically significant change in fuel efficiency for the tested vehicle model.

Best Practices for Reporting and Validity

Achieving a statistically significant result is only half the process; researchers must ensure their findings are reported comprehensively and accurately. A complete statistical summary for the Wilcoxon Signed-Rank Test must include several mandatory components: the test name, the total sample size (N=12), the derived test statistic (W = 10.5), and the associated p-value (p = 0.044). Crucially, because this is a non-parametric procedure, the descriptive statistics provided should emphasize the **median** and the **interquartile range (IQR)**, as these metrics are more appropriate indicators of central tendency and variability than the mean and standard deviation.

Beyond merely establishing statistical significance, it is strongly recommended practice to calculate and report an appropriate measure of effect size. The effect size quantifies the practical magnitude

of the observed difference, offering a critical complement to the p-value. While the p-value informs us about the probability of the difference occurring by chance, the effect size tells us **how large** that difference actually is. A scenario where a small p-value coexists with a negligible effect size suggests a statistically significant finding that may be practically unimportant or trivial in a real-world context. Common measures of effect size for this test include r (often derived from the Z-score) or Kendall's W.

Finally, maintaining the validity of your analytical conclusions requires continuous confirmation that the initial assumption of a **paired data structure** remains accurate throughout the research design. If, hypothetically, your research involved two entirely separate, independent groups (e.g., Group A receiving the treatment and Group B receiving a placebo, where the subjects in A and B are different individuals), applying the Wilcoxon Signed-Rank Test would be inappropriate and invalid. In such a case, the correct non-parametric analytical methodology would be the Mann-Whitney U Test. Selecting the precise statistical test that aligns with the experimental design is paramount for ensuring the integrity and trustworthiness of the research findings.