Learn How to Perform a Mann-Whitney U Test in Python

Authored by

Mohammed looti

November 8, 2025

The Mann-Whitney U Test, often recognized as the Wilcoxon rank-sum test, stands as a cornerstone in non-parametric statistics, serving as a critical tool for comparing the distributions of two independent samples. Its primary utility lies in assessing whether one sample tends to have values stochastically larger or smaller than the other, effectively testing for differences in location (median). This test is indispensable in modern data analysis, particularly when fundamental assumptions required by traditional parametric procedures cannot be satisfied, such as the crucial requirement of data normality or when sample sizes are small, typically defined as having fewer than 30 observations per group. Mastering the appropriate application of this test is the initial step toward achieving robust and reliable data analysis when working with real-world datasets that frequently defy ideal distributions.

The strength of the Mann-Whitney U test derives from its unique methodology, which involves ranking all observations from both samples combined, rather than relying on the raw data values themselves. By evaluating the probability that a randomly selected observation from the first group will exceed a randomly selected observation from the second group, the test effectively assesses differences in central tendency without assuming a specific distribution shape. This versatility makes it the most powerful and widely accepted nonparametric alternative to the independent samples T-test, which is constrained by strict assumptions regarding normality and homogeneity of variance. This comprehensive tutorial provides a systematic, step-by-step guide detailing how to efficiently execute and accurately interpret the results of a Mann-Whitney U test utilizing the potent statistical capabilities embedded within the Python programming environment.

Understanding the Mann-Whitney U Test

The essential goal of the Mann-Whitney U Test is to assess potential differences between two population medians based on data collected from independent samples. Unlike parametric tests, which focus strictly on comparing means and necessitate interval or ratio data that adhere to a specific distribution (most commonly the normal distribution), the U test operates exclusively on the ranks of the data. This focus on ordered ranking provides intrinsic resistance to the distorting effects of outliers and significant skewness, conditions that can severely compromise the statistical validity of results obtained from parametric tests. When dealing with empirical data across diverse fields--be it clinical trials, social science experiments, or quality control metrics--it is extremely common to encounter datasets that naturally violate the normality assumption, making the Mann-Whitney U Test a necessary and statistically sound choice for comparison.

The computational process begins by pooling all observations from the two distinct groups and assigning them a combined rank, ordered sequentially from the smallest observation (rank 1) to the largest. Following this comprehensive ranking, the test calculates the sum of the ranks for each group individually. The resulting U statistic quantifies the degree of overlap between the rank distributions of the two groups. If both samples are drawn from identical populations, statistical

expectation dictates that their rank sums should be approximately equivalent, yielding a U statistic close to its theoretical mean. Conversely, a U statistic that is significantly large or small suggests a genuine disparity, indicating that the two groups were likely drawn from populations with different central tendencies. This reliance on the ordinal position (ranks) rather than the absolute magnitude of the raw data values is the fundamental characteristic that defines its nonparametric methodology.

Selecting the statistically appropriate test is critical for ensuring that any conclusions drawn are accurate and defensible. If the collected data rigorously satisfies the prerequisites for parametric testing--specifically normality, interval data measurement, and adequate sample size--the T-test is generally the preferred option because it possesses slightly greater statistical power. However, the misapplication of a T-test to data that is markedly non-normal, particularly in scenarios involving small samples, dramatically increases the likelihood of committing either a Type I error (false positive) or a Type II error (false negative). Therefore, rigorous preliminary data checks, often involving procedures such as the Shapiro-Wilk test for normality or visual aids like Q-Q plots, are mandatory. If these checks reveal a significant deviation from the normal distribution, the Mann-Whitney U test serves as the responsible, statistically robust alternative, preserving the integrity and reliability of the inferential process.

Case Study: Evaluating Fuel Treatment Effectiveness

To provide a clear demonstration of the Mann-Whitney U test's practical application, let us consider a typical research scenario within the automotive sector. A research team is dedicated to evaluating the efficacy of a newly formulated, proprietary fuel treatment designed to enhance overall vehicle performance and fuel efficiency. Their primary research objective is to rigorously determine whether the application of this treatment yields a quantifiable, statistically significant change in the average miles per gallon (mpg) achieved by vehicles operating under identical, controlled testing conditions. The experimental design mandates the use of two independent groups of vehicles: an experimental group that receives the new fuel treatment and a control group that receives no such additive.

The researchers successfully collected relevant data from 12 vehicles assigned to the treatment group and 12 vehicles assigned to the control group. Given that the sample size for each group is quite small ($n = 12$), falling substantially below the $n > 30$ threshold commonly relied upon to invoke the Central Limit Theorem, and acknowledging that external variables often introduce skewness into fuel economy measurements, the researchers determined that assuming a normal distribution for the mpg values would be statistically inappropriate and high-risk. Consequently, they judiciously chose to employ the Mann-Whitney U Test. This strategic choice allows them to test rigorously for a statistically significant difference in median mpg between the two groups without imposing restrictive distributional assumptions, thereby ensuring their final conclusions are

robust despite the inherent limitations associated with a small sample size.

Prior to initiating the computational analysis, it is essential to formally establish the precise hypotheses that the statistical test is designed to address. The core research question must be translated directly into a defined set of statistical hypotheses. The **Null Hypothesis (H0)** formally postulates that there is absolutely no statistical difference in the median miles per gallon achieved between the cars receiving the fuel treatment and those in the control group. Conversely, the **Alternative Hypothesis (HA)** asserts that a statistically significant difference does exist in the median mpg measurements between the two groups. By executing the U test, the goal is to gather evidence strong enough to warrant the rejection of H0 in favor of HA, which would constitute a statistical demonstration that the fuel treatment successfully produced a measurable impact on vehicle fuel efficiency.

Step 1: Preparing the Data in Python

The foundational step in conducting any rigorous statistical analysis using Python involves organizing the raw observational data into the requisite data structures. For the Mann-Whitney U test, this requires the creation of two distinct arrays or lists, one designated for each independent group of observations. In the context of our fuel treatment case study, these arrays will store the miles per gallon (mpg) measurements corresponding to the treated cars and the untreated cars, respectively. For this demonstration, we will utilize standard Python lists to efficiently represent these datasets, which will then be passed as arguments directly to the chosen statistical function.

The following code block meticulously defines the two datasets, clearly identifying `group1` as the Treatment group and `group2` as the Control group. It is considered best practice to verify that the data has been transcribed accurately and that both arrays contain the correct number of observations, although the Mann-Whitney U test does not strictly require the groups to have equal sample sizes, unlike certain other comparative statistical tests. This definition of the data structure is an absolute prerequisite for importing and subsequently calling the necessary functions from the specialized statistical libraries, thereby laying the essential computational groundwork for the analytical phase that follows.

group1 =
group2 =

It is important to acknowledge that while simple native Python lists are sufficient for this example involving small, straightforward samples, for handling larger, more intricate datasets in professional settings, it is both more common and significantly more efficient to leverage data structures provided by the powerful **NumPy** library or the robust **Pandas DataFrame** structure. These specialized libraries are designed to offer optimized array handling, vectorization, and advanced

data manipulation capabilities that are indispensable tools for serious data science work in Python. Nonetheless, for a simple, direct comparison of two small independent samples, the native list structure adequately serves the purpose of interfacing the data with the statistical functions provided by the Scipy library.

Step 2: Executing the Test using Scipy's mannwhitneyu

With the sample data properly prepared and structured, the subsequent crucial phase involves the computational execution of the Mann-Whitney U test itself. Python's extensive scientific computing ecosystem offers exceptional statistical functionality, primarily channeled through the Scipy library, specifically via the `scipy.stats` module. This module conveniently hosts the dedicated function required for our analysis: `mannwhitneyu()`. This function is engineered to efficiently perform all the necessary internal rank-sum calculations and subsequently returns the computed U statistic alongside the crucial associated p-value.

To ensure the accurate utilization of this powerful function, it is essential to fully grasp its core syntax and operational parameters, as minor adjustments in configuration can lead to fundamentally different interpretations of the final results. The `mannwhitneyu()` function adheres to the following flexible structure, allowing the analyst to define precisely how the test is applied to the data:

mannwhitneyu(x, y, use_continuity=True, alternative=None)

The primary arguments dictate the data input and the test configuration:

x: This parameter strictly represents the array of sample observations originating from the first independent group (designated as `group1` in our case study).
y: This parameter represents the array of sample observations from the second independent group (designated as `group2`).
use_continuity: This is a boolean flag that determines whether a continuity correction (specifically of 1/2) should be applied during calculation. This correction is conventionally set to **True** by default, particularly when working with small sample sizes, as it helps improve the approximation of the discrete rank sum distribution using a continuous normal distribution.
alternative: This critical parameter defines the directional nature of the test being conducted, which directly corresponds to the formulation of the alternative hypothesis (HA). The available options are:

`None` (Default): Computes a p-value that is exactly half the size of the `two-sided` p-value.
`'two-sided'`: Utilized when the hypothesis tests simply whether the two distributions are statistically different (HA: Group 1 neq Group 2).
`'less'`: Used for a one-sided test where the specific hypothesis is that Group 1's median is

statistically less than Group 2's median.
`'greater'`: Used for a one-sided test where the specific hypothesis is that Group 1's median is statistically greater than Group 2's median.

In the context of our fuel treatment example, the research team is interested solely in determining whether the median mpg is different between the two groups, without hypothesizing a specific direction (i.e., whether it increased or decreased). Therefore, a **two-sided** test is the most appropriate and conservative choice to capture any statistically significant deviation from the null hypothesis. The resulting Python code below imports the necessary statistical library and executes the test using the defined parameters, generating the precise statistical output required for final interpretation.

import scipy.stats as stats

#perform the Mann-Whitney U test
stats.mannwhitneyu(group1, group2, alternative='two-sided')

(statistic=50.0, pvalue=0.2114)

Step 3: Interpreting the Statistical Output

The successful execution of the `mannwhitneyu` function generates two essential numerical values: the calculated U statistic and the resulting p-value. In the specific context of our case study, the output generated is `(statistic=50.0, pvalue=0.2114)`. The interpretation of these two critical statistics dictates the final conclusion regarding the claimed effectiveness of the experimental fuel treatment.

We must recall the formally established hypotheses for this two-sided test, which guide our decision-making process:

H0: The median miles per gallon is statistically equal between the fuel treatment group and the control group.

HA: The median miles per gallon is *not* statistically equal between the two groups.

To arrive at a statistical decision, the calculated p-value must be compared against a pre-determined significance level, conventionally denoted by the Greek letter alpha ($alpha$). Standard scientific practice sets this significance level at $alpha = 0.05$. This value explicitly represents the maximum acceptable probability of erroneously rejecting the **Null Hypothesis** when it is, in reality, true--a statistical mistake known as a Type I error.

The decision rule is unambiguous and straightforward: if the calculated p-value is strictly less than

$alpha$ (0.05), we possess sufficient evidence to reject the **Null Hypothesis** (H0). Conversely, if the p-value is greater than or equal to $alpha$, we must formally fail to reject H0. In this particular instance, our computed p-value is **0.2114**. Since 0.2114 is substantially greater than the threshold of 0.05, the statistical imperative requires us to fail to reject the **Null Hypothesis**.

The definitive statistical conclusion is that, based on the results of the Mann-Whitney U Test, there is inadequate evidence to assert a statistically significant difference in the median mpg between the vehicles treated with the new fuel additive and those that received no treatment. In terms of practical application, while the raw sample data might show minor numerical variations, these observed differences are statistically attributed to random sampling variation rather than a genuine, measurable effect of the treatment itself. Consequently, the researchers cannot confidently conclude that the fuel treatment successfully or significantly altered the cars' fuel efficiency.

Additional Resources and Further Study

While the Python implementation, leveraging the efficiency and power of the Scipy library, offers a streamlined approach to executing the Mann-Whitney U test, developing a comprehensive understanding of how this test is applied across various statistical environments is highly beneficial for the well-rounded data scientist. The fundamental underlying statistical theory remains universally consistent, even though the specific syntax, function names, and output presentation can vary significantly across different software packages and programming languages.

The following resources are recommended to help users explore the application and conceptual nuances of the Mann-Whitney U Test using other popular statistical software and programming tools, which may be necessitated depending on specific organizational, academic, or industry standards:

Performing the Mann-Whitney U Test in R (Typically executed using the versatile `wilcox.test()` function).
Executing the Test in SPSS (Generally located under the menu path: Nonparametric Tests > Legacy Dialogs > 2 Independent Samples).
Detailed manual calculation of the U statistic for didactic purposes, focusing on a deep understanding of the rank-sum methodology.

Ultimately, achieving mastery of the Mann-Whitney U Test equips analysts with the necessary confidence and expertise to handle real-world data that violates the restrictive assumptions of parametric statistics. This mastery is crucial for maintaining analytical rigor and ensuring the highest reliability in statistical conclusions, irrespective of the inherent complexity of the dataset or the specific domain of research being investigated.