How to Perform a Kruskal-Wallis Test in Python

Authored by

stats writer

December 25, 2025

The field of data analysis often requires researchers to compare multiple independent groups to determine if they originate from the same distribution. When dealing with data that fails to meet the strict requirements of traditional parametric tests--such as Analysis of Variance (ANOVA)--statisticians turn to powerful alternatives. One of the most vital tools in this arsenal is the Kruskal-Wallis Test, often referred to as the Kruskal-Wallis H Test. This robust procedure allows us to assess whether there are statistically significant differences among the medians of three or more independent samples, particularly when the underlying assumptions of normality and homogeneity of variances are violated.

The Kruskal-Wallis Test is classified as a non-parametric statistical test, meaning it does not rely on assumptions about the specific shape of the population distribution. Instead of comparing means, as parametric tests do, it compares the ranks of the observations across all groups. This makes it an incredibly versatile and reliable choice for analyzing ordinal data or continuous data where the distribution is heavily skewed or contains significant outliers. Understanding how and when to apply this test is crucial for generating accurate, defensible statistical conclusions in diverse research areas, from social sciences to biological studies.

This comprehensive guide will detail the theoretical foundation of the Kruskal-Wallis Test and provide a step-by-step tutorial on how to implement and interpret it effectively using the powerful statistical capabilities available in the Python programming language, specifically leveraging the functionalities provided by the scipy.stats library. By the end of this tutorial, you will be equipped to perform robust comparisons among multiple independent samples in your own data analysis projects.

The Necessity of Non-Parametric Testing

In classical statistical inference, tests like the T-test or ANOVA operate under strict conditions. Primarily, they assume that the dependent variable follows a normal distribution within each group and that the variances across these groups are approximately equal (homoscedasticity). While these assumptions often hold true for large, well-behaved datasets, real-world data frequently defies these constraints. Datasets derived from surveys, skewed observational studies, or experiments with small sample sizes often exhibit non-normal distributions, making the use of parametric tests inappropriate and potentially leading to incorrect p-value calculations and flawed conclusions.

When the assumption of normality is violated, using a test based on ranks becomes essential. The Kruskal-Wallis Test provides a safe and reliable alternative in these scenarios. It is fundamentally an extension of the Wilcoxon Rank-Sum Test (Mann-Whitney U Test) for situations involving two groups, generalizing the concept to three or more groups. By utilizing the ranks of the data rather than the raw data values themselves, the test minimizes the influence of extreme outliers and

allows the researcher to focus on whether the distributions are generally shifted relative to one another, which is typically interpreted as a difference in medians.

It is crucial to understand that while the Kruskal-Wallis Test compares medians, it is technically testing the **null hypothesis** that the samples are drawn from the same population or populations with identical shapes and scales. If the distributions are known to have similar shapes but differ only in location, then rejecting the null hypothesis definitively implies a difference in medians. Researchers must consider the underlying data distribution carefully before interpreting the results solely based on median differences.

Understanding the Kruskal-Wallis H Test Statistic

The Kruskal-Wallis Test calculates a test statistic, H, which measures the extent to which the average ranks for the different groups vary. The procedure involves pooling all observations from all groups together and ranking them from smallest to largest. Ties are handled by assigning the average of the ranks they would have received. Once ranked, the sum of the ranks for each group is calculated. The H statistic quantifies how much these observed rank sums deviate from the rank sums that would be expected if the null hypothesis were true (i.e., if all groups were truly identical).

The formula for the H statistic is complex, involving the total number of observations, the sample size of each group, and the sum of the ranks within each group. Mathematically, larger values of H indicate greater disparities among the group ranks, suggesting that the groups are indeed different. When the sample sizes are sufficiently large (generally five or more observations per group), the distribution of the H statistic can be closely approximated by a Chi-Square distribution with $k-1$ degrees of freedom, where k is the number of groups being compared. This approximation is what allows us to calculate the critical value and the associated p-value necessary for hypothesis testing.

This reliance on the Chi-Square distribution for calculating the p-value is a key characteristic of the Kruskal-Wallis Test, enabling statistical software like Python's scipy.stats module to quickly determine the probability of observing the calculated H statistic if the **null hypothesis** were true. Understanding the statistical theory behind the H statistic helps solidify the interpretation process, ensuring that the decision to reject or fail to reject the null hypothesis is grounded in sound statistical principles rather than just a simple cutoff rule.

Key Assumptions for the Kruskal-Wallis Test

Although the Kruskal-Wallis Test is non-parametric and therefore avoids the restrictive assumptions of normality and homoscedasticity, it is not entirely assumption-free. To ensure the validity and proper interpretation of the results, several key assumptions must still be met. Firstly,

the samples must be **independent**. This means that the observations in one group cannot influence or be related to the observations in any other group. For instance, in an experiment comparing fertilizer effects, the growth of a plant in group 1 must not be dependent on the growth of a plant in group 2.

Secondly, the variable of interest must be measurable on at least an ordinal scale. This means the data must be capable of being ranked. Since the test fundamentally relies on converting raw scores into ranks, if the data cannot be ordered, the test cannot be performed. Fortunately, most continuous or discrete quantitative measurements meet this requirement easily. Thirdly, and perhaps most critically for accurate interpretation, it is generally assumed that the distributions of the populations being compared have the same shape. If the distributions have drastically different shapes (e.g., one is heavily skewed and the other is symmetric), rejecting the **null hypothesis** might indicate differences in shape or spread, rather than just differences in location (median).

Finally, researchers must ensure that the observations are selected randomly from their respective populations. **Random sampling** is a foundational requirement for almost all inferential statistics, guaranteeing that the samples are representative of the larger populations they are intended to model. Adherence to these assumptions ensures that the calculated p-value accurately reflects the probability of observing the data under the null hypothesis, thus strengthening the reliability of the overall statistical conclusion.

Case Study: Investigating Plant Growth and Fertilizers

To illustrate the practical application of the Kruskal-Wallis Test, let us consider a common experimental scenario. Imagine agricultural researchers are investigating the efficacy of three different fertilizers--call them Fertilizer A, Fertilizer B, and Fertilizer C--on promoting plant growth. They hypothesize that at least one of these fertilizers leads to a significantly different median plant height after a fixed growth period. They decide on a robust experimental design, randomly selecting 30 genetically similar plants and dividing them equally into three independent groups of 10. Each group receives a specific fertilizer type, and all other environmental variables (light, water, soil type) are meticulously controlled to isolate the effect of the treatment.

After one full month of treatment, the researchers carefully measure the height (in centimeters) of every single plant. Since initial exploratory data analysis suggests that the height measurements might not follow a perfect normal distribution--perhaps due to a few plants experiencing stunted or unusually accelerated growth--the parametric assumption of ANOVA might be violated. Consequently, the Kruskal-Wallis Test is the statistically appropriate tool to compare the median growth across the three fertilizer groups. The core question they seek to answer is: Is the median plant growth statistically the same across all three fertilizer treatments?

The formal hypotheses driving this analysis are clearly defined. The **null hypothesis (H_0)**

states that the median plant heights for all three fertilizer groups are equal. Conversely, the alternative hypothesis (H_a) posits that at least one group's median plant height is different from the others. The research objective is to use the resulting p-value from the statistical test to determine whether there is sufficient evidence to reject the null hypothesis in favor of the alternative, thereby concluding that the type of fertilizer has a significant impact on plant growth.

Setting Up the Python Environment and Data

Performing statistical analysis in Python requires importing specialized libraries that contain the necessary functions. For the Kruskal-Wallis Test, the primary tool we rely upon is the statistical sub-module within the SciPy library, specifically scipy.stats. Before running the test, the first necessary step is to structure the collected data appropriately. In Python, it is standard practice to represent independent samples as distinct lists or arrays, making them easily accessible for the statistical function.

Following the case study example, we input the height measurements collected from the 30 plants. Each list corresponds precisely to one treatment group. This step involves meticulous data entry to ensure accuracy, as any error here would invalidate the subsequent statistical analysis. Below, we define the three data arrays representing the heights (in cm) achieved by plants using Fertilizer A (group1), Fertilizer B (group2), and Fertilizer C (group3). This organization is essential as the `kruskal` function expects each group to be passed as an independent argument.

group1 =
group2 =
group3 =

Once the data is correctly entered into these variables, the Python environment is ready for the computational step. Although this example uses simple lists for demonstration, in real-world scenarios, data is often loaded from external files (like CSVs) using libraries such as Pandas, which would then require extracting the relevant columns into separate arrays before passing them to the **scipy.stats** function.

Executing the Kruskal-Wallis Test in Python

The efficiency of Python for statistical computation is largely due to comprehensive libraries like SciPy. To perform the Kruskal-Wallis Test, we must import the `stats` module and then call the specific function designed for this purpose: `stats.kruskal()`. This function is designed to handle multiple input arrays simultaneously, which perfectly suits the requirements of a three-or-more-group comparison. The syntax is clean and intuitive, requiring only the names of the data arrays as arguments.

When executing `stats.kruskal(group1, group2, group3)`, the function internally performs all the necessary steps: pooling the data, assigning ranks to every observation, handling ties, calculating the H statistic, and finally, using the Chi-Square approximation to derive the corresponding p-value. The function returns a tuple containing two key metrics: the calculated H test statistic and the two-sided **p-value**.

Executing the code for our plant growth data yields the following output, which encapsulates the entire statistical comparison:

from scipy import stats

#perform Kruskal-Wallis Test
stats.kruskal(group1, group2, group3)

(statistic=6.2878, pvalue=0.0431)

This output provides us with the necessary numerical evidence to proceed to the crucial stage of interpretation. The test statistic is calculated as 6.2878, and the associated probability, the p-value, is 0.0431. These two numbers are the foundation upon which we make our statistical decision regarding the effect of the different fertilizers.

Interpreting the Test Output and P-Value

Interpretation of the Kruskal-Wallis Test results follows the standard procedure of hypothesis testing. We must compare the calculated p-value against a pre-determined significance level ($alpha$). Most commonly, researchers use an alpha level of $alpha = 0.05$. This threshold represents the maximum acceptable risk of making a Type I error--rejecting the **null hypothesis** when it is, in fact, true.

The hypotheses established for our fertilizer study were:

The null hypothesis (H_0): The median plant height is equal across all three fertilizer groups.

The alternative hypothesis (H_a): The median plant height is not equal across all three fertilizer groups (at least one median differs).

In our example, the test statistic is $H = strong >6.2878$ and the corresponding **p-value** is $strong >0.0431$. Since 0.0431 is less than the conventional significance level of 0.05, we meet the criterion for statistical significance. Therefore, we **reject the null hypothesis** (H_0). Rejecting the null hypothesis means we have found sufficient statistical evidence to conclude that the distribution of plant growth is not the same across all three fertilizer groups; practically, this suggests that at least one fertilizer leads to a different median plant height compared to the others.

Post-Hoc Analysis: When H0 is Rejected

A crucial limitation of the Kruskal-Wallis Test is that, much like ANOVA, it is an omnibus test. If the **null hypothesis** is rejected, it merely tells us that a difference exists somewhere among the groups, but it does not specify which pairs of groups are significantly different from each other (e.g., is Fertilizer A different from B, or B from C?). To pinpoint these specific differences, a follow-up procedure known as a post-hoc test is required.

When using the Kruskal-Wallis Test, the appropriate non-parametric post-hoc test involves performing multiple pairwise comparisons using a method like the Dunn's Test or the Conover-Iman test, and critically, applying a correction for multiple comparisons. Performing simple pairwise Mann-Whitney U tests without correction would inflate the overall Type I error rate. Common correction methods used alongside these non-parametric comparisons include the Bonferroni correction or the Benjamini-Hochberg procedure, which adjust the individual p-values to maintain the family-wise error rate at the desired $alpha$ level.

While the standard scipy.stats library provides the core Kruskal-Wallis functionality, specialized libraries, such as `scikit-posthocs` in Python, are often necessary to perform the non-parametric post-hoc tests accurately with the necessary corrections. Researchers should always ensure they follow up a significant Kruskal-Wallis result with an appropriate post-hoc analysis to provide a complete and actionable statistical report, identifying the source of the statistically significant difference.

Conclusion and Best Practices

The Kruskal-Wallis Test remains an indispensable tool for researchers dealing with multiple independent samples, particularly when the data violates the assumptions required for parametric methods. Its rank-based approach provides a robust measure of location differences (medians) without relying on assumptions of normality, making it highly versatile for real-world data analysis, especially within the Python ecosystem facilitated by scipy.stats.

To ensure best practices when performing this test, analysts should always start with thorough exploratory data analysis (EDA), including visualizing the distributions (e.g., using box plots) and assessing potential skewness or outliers. This initial check confirms the necessity of using a non-parametric statistical test. Furthermore, always clearly state the **null hypothesis** and the chosen significance level ($alpha$) before executing the test, preventing post-hoc bias in decision-making.

By following the steps outlined in this tutorial--entering data, executing the `stats.kruskal()` function, and interpreting the resulting test statistic and p-value--you can confidently determine whether differences exist among your groups. Remember, rejecting the null hypothesis only opens the door to further investigation, requiring dedicated post-hoc analysis to fully understand the

pairwise relationships driving the overall statistical significance.