

How to Perform a Kruskal-Wallis Test: A Step-by-Step Guide with Example

Authored by
stats writer

March 1, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Perform a Kruskal-Wallis Test: A Step-by-Step Guide with Example*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=133338>

Understanding the Fundamentals of the Kruskal-Wallis Test

The **Kruskal-Wallis test**, often referred to as the Kruskal-Wallis H test, serves as a cornerstone in the field of **non-parametric statistics**. This robust analytical tool is specifically designed to evaluate whether there are statistically significant differences between three or more independent groups. Unlike parametric tests that rely on specific distributional parameters, the Kruskal-Wallis test operates by analyzing the **medians** of the groups rather than their means, making it an essential instrument for researchers dealing with data that does not conform to standard assumptions of normality.

In many scientific investigations, researchers find themselves comparing multiple treatment groups or conditions where the underlying data is measured on an **ordinal scale** or where the sample sizes are too small to justify the assumption of a bell-shaped curve. In such scenarios, the **Kruskal-Wallis test** provides a reliable alternative to the traditional analysis of variance. By transforming raw scores into ranks, the test effectively minimizes the influence of extreme values, thereby offering a more accurate reflection of the central tendency within skewed datasets. This rank-based approach ensures that the analysis remains valid even when the data exhibits high levels of **skewness** or kurtosis.

The versatility of the **Kruskal-Wallis test** extends across various disciplines, including psychology, biology, and market research. For instance, a clinical psychologist might use this test to compare the efficacy of four different therapeutic interventions on patient anxiety levels, or an ecologist might use it to assess plant growth across different soil types. Because the test does not require the assumption of a normal distribution, it allows for a broader range of variables to be analyzed with statistical rigor. Consequently, understanding how to implement and interpret this test is vital for any researcher aiming to produce high-quality, reproducible scientific findings.

Furthermore, it is important to recognize that the **Kruskal-Wallis test** is an "omnibus" test. This means that while it can tell a researcher that at least one group is different from the others, it does not specify which particular groups differ. If the test yields a significant result, further post-hoc testing is required to pinpoint the exact nature of the differences. This multi-stage approach ensures that researchers do not make unfounded claims about their data but instead follow a systematic path toward discovering the true relationships within their experimental observations.

Distinguishing Between Parametric and Non-Parametric Methods

When selecting a statistical method, researchers must often choose between parametric and non-parametric pathways. The **one-way ANOVA** is the parametric counterpart to the **Kruskal-Wallis test**. While the ANOVA is highly powerful, its validity is strictly contingent upon several rigorous assumptions, most notably that the residuals of the model are normally distributed and that the

groups exhibit **homoscedasticity**, or equal variance. When these conditions are met, the ANOVA is the preferred method; however, real-world data frequently fails to meet these criteria, necessitating the use of the **Kruskal-Wallis test**.

The primary advantage of the **Kruskal-Wallis test** over the ANOVA lies in its resilience to **outliers**. In a parametric ANOVA, a single extreme value can significantly pull the mean and inflate the variance, potentially leading to a Type I or Type II error. In contrast, the **Kruskal-Wallis test** converts these extreme values into ranks. Since an outlier's rank is only one step away from the next highest value, its mathematical weight in the final calculation is greatly reduced. This makes the non-parametric approach far more stable in the presence of noisy or contaminated data.

Another critical distinction is the level of measurement required for each test. The **one-way ANOVA** generally requires a **continuous variable** measured on an interval or ratio scale. The **Kruskal-Wallis test**, however, is perfectly suited for an **Likert Scale** or other ordinal data where the distance between units is not necessarily uniform. For example, in a survey where respondents rate their satisfaction from 1 to 5, the difference between "neutral" and "satisfied" may not be the same as the difference between "satisfied" and "very satisfied." The rank-based logic of the Kruskal-Wallis method respects this limitation of the data.

Finally, researchers often turn to the **Kruskal-Wallis test** when dealing with small sample sizes. While the **central limit theorem** suggests that large samples will approximate normality, small samples provide no such guarantee. By using a non-parametric approach, the researcher avoids making unjustified assumptions about the population distribution. This cautious approach to data analysis is a hallmark of rigorous scientific inquiry, ensuring that the conclusions drawn are supported by the actual evidence rather than mathematical artifacts of an inappropriate model.

Core Assumptions and Data Requirements

Despite being a non-parametric method, the **Kruskal-Wallis test** is not entirely assumption-free. To ensure the validity of the results, researchers must adhere to several key requirements. First and foremost is the assumption of **independence**. This means that the observations in each group must be independent of one another, and there must be no relationship between the members of different groups. In practice, this is usually achieved through random sampling and ensuring that the groups represent distinct, non-overlapping populations.

The second major requirement concerns the nature of the **dependent variable**. The response variable under investigation should be at least ordinal or continuous. If the data consists of nominal categories with no inherent order, the **Kruskal-Wallis test** cannot be applied. Examples of appropriate data include test scores, weights, or rankings. By ensuring that the data possesses a meaningful order, the ranking process remains logically consistent and mathematically sound throughout the calculation of the test statistic.

A third, and often overlooked, assumption is that the distributions of the groups must have a similar shape. While the **Kruskal-Wallis test** does not require the distributions to be normal, it does require them to be comparable in terms of their **dispersion** and overall profile. If one group is highly skewed to the left while another is skewed to the right, the test may detect a difference in distribution rather than a specific difference in **medians**. Therefore, visualizing the data through boxplots or histograms prior to testing is a highly recommended practice.

Moreover, the test assumes that the groups are independent and not "matched" or "repeated measures." If a researcher were measuring the same group of individuals at three different time points, the **Kruskal-Wallis test** would be inappropriate. In that specific context, the **Friedman test** would be the correct non-parametric choice. Maintaining clarity on the experimental design is essential to choosing the correct statistical framework and avoiding the common pitfalls associated with data analysis.

The Mathematical Architecture of the H-Statistic

The calculation of the **Kruskal-Wallis test** centers on the generation of the H-statistic. This value represents the degree of discrepancy between the observed ranks of the groups and the ranks that would be expected if the **null hypothesis** were true. The process begins by pooling all observations from every group into a single dataset and ranking them from smallest to largest. In the event of ties, where two or more observations have the same value, each is assigned the average of the ranks they would have otherwise occupied.

The formal **formula** for the H-statistic is expressed as $H = \frac{12}{N(N+1)} \sum (R_i^2 / n_i) - 3(N+1)$. In this equation, N represents the total number of observations across all groups, n_i denotes the number of observations in a specific group, and R_i is the sum of the ranks for that group. This formula effectively measures the variance of the rank sums; a higher H-value suggests that the ranks are not distributed evenly across the groups, pointing toward a significant difference in the population **medians**.

To understand the intuition behind this formula, consider that if all groups were identical, their average ranks would be nearly equal. The term $\sum (R_i^2 / n_i)$ would be relatively small, leading to a low H-statistic. Conversely, if one group consistently occupies the highest ranks and another the lowest, the squared sums of those ranks will be much larger, increasing the value of H. This mathematical structure allows the test to capture the magnitude of separation between the groups while remaining anchored in the simple logic of ordinal ranking.

Once the H-statistic is calculated, it must be evaluated against a **chi-square distribution**. This is because, as the sample size increases, the distribution of the H-statistic approximates the chi-square distribution with $k-1$ **degrees of freedom**, where k is the number of groups. This approximation allows researchers to determine the **p-value**, which quantifies the probability of

observing such a result by chance. If the p-value is below a predetermined threshold, typically 0.05, the result is considered statistically significant.

A Detailed Walkthrough of the Ranking Process

To truly grasp the mechanics of the **Kruskal-Wallis test**, it is helpful to walk through the ranking process using a concrete example. Imagine a scenario where a researcher is comparing the performance of three different store branches (Store A, Store B, and Store C) based on customer satisfaction scores. The first step is to take every single score from all three stores and place them in a single ordered list. This unified ranking is crucial because it establishes a common baseline for comparison across the independent groups.

If Store A has scores of 8 and 9, and Store B has scores of 5 and 6, the ranking would proceed by assigning the lowest overall score (let's say 4 from Store C) the rank of 1. The next highest score gets rank 2, and so on, until the highest score in the entire dataset receives the final rank of N. This transformation from raw scores to ranks is what grants the **Kruskal-Wallis test** its **non-parametric** power, as it focuses on the relative position of data points rather than their absolute magnitude.

After the universal ranking is complete, the researcher separates the ranks back into their original groups. For Store A, the ranks might be 11, 13, 15, and so on. The sum of these ranks (R_i) is then calculated for each store. These rank sums are the primary inputs for the H-statistic formula. If the sum of ranks for Store A is significantly higher than those for Store B or C, it indicates that Store A's customers generally reported higher satisfaction levels, even if the distribution of those scores was not normal.

One must also account for tied ranks, which occur when multiple observations share the same value. In a dataset of 15 observations, if the 4th and 5th values are both "7", they are both assigned a rank of 4.5. While a large number of ties can require a correction factor in the H-statistic formula, most modern statistical software packages handle this automatically. The goal remains the same: to create a mathematical representation of how the groups are positioned relative to one another in the overall data landscape.

Statistical Significance and the Chi-Square Distribution

The final stage of the **Kruskal-Wallis test** involves determining the significance of the calculated H-statistic. This is done by comparing the H-value to a critical value from the **chi-square distribution** table. The **degrees of freedom** for this comparison are determined by the number of groups being compared minus one ($k - 1$). For example, if a study compares three different drugs, the degrees of freedom would be 2. If the calculated H is larger than the critical value found in the table for that specific level of **statistical significance**, the null hypothesis is rejected.

The use of the chi-square distribution is based on the assumption that the sample sizes are sufficiently large (usually $n > 5$ per group). When sample sizes are very small, the chi-square approximation may not be as accurate, and researchers may need to consult exact probability tables specifically designed for the **Kruskal-Wallis test**. However, in the vast majority of modern research applications, the chi-square approximation is considered a standard and reliable method for determining the **p-value**.

A significant result ($p < 0.05$) indicates that there is a less than 5% probability that the observed differences in group **medians** occurred due to random chance. In such cases, the researcher concludes that the independent variable (e.g., the type of drug or the store location) has a measurable effect on the dependent variable. It is important to remember that significance does not imply "importance" or "magnitude of effect," but rather the reliability of the observed difference.

If the H-statistic does not exceed the critical value, the researcher fails to reject the **null hypothesis**. This suggests that any observed differences between the groups are likely due to sampling error or random variation. In this situation, it cannot be concluded that the groups are different. Failing to reject the null is a common outcome in scientific research and serves as a necessary check against over-interpreting data that does not provide strong evidence for an effect.

Practical Application: Analyzing Drug Efficacy for Knee Pain

To illustrate the application of these concepts in a professional research setting, let us examine a study involving three different medications for knee pain. A researcher recruits 30 participants, all suffering from similar levels of joint discomfort. These individuals are randomly assigned to three groups, with each group receiving a different treatment: Drug 1, Drug 2, or Drug 3. After a month-long trial, participants rate their pain on a scale of 1 to 100, where higher scores indicate more intense pain. This type of **ordinal variable** is a perfect candidate for the **Kruskal-Wallis test**.

The data collected from the 30 individuals is presented in the following table:

Drug 1	Drug 2	Drug 3
78	71	57
65	66	88
63	56	58
44	40	78
50	55	65
78	31	61
70	45	62

Drug 1	Drug 2	Drug 3
61	66	44
50	47	48
44	42	77

The researcher's objective is to determine if the **significance level** of 0.05 will reveal a statistically meaningful difference between these treatments. Given the nature of the pain scale and the potential for **outliers** or non-normal distributions in self-reported pain scores, the **Kruskal-Wallis test** is the most appropriate choice. This ensures that the analysis is not skewed by any one individual's extreme reaction to a medication.

The analysis follows a structured protocol: stating the hypotheses, performing the ranking and H-calculation, and interpreting the results. By using a specialized **statistical software** or an online calculator, the researcher can efficiently process the data. This systematic approach is vital for maintaining the integrity of the clinical trial and ensuring that any medical recommendations based on the study are grounded in sound statistical evidence.

Navigating the Null and Alternative Hypotheses

Every **Kruskal-Wallis test** begins with a clear statement of the hypotheses. The **null hypothesis** (H_0) acts as the default position, asserting that there is no difference between the groups. Specifically, it states that the population **medians** of the knee-pain ratings for Drug 1, Drug 2, and Drug 3 are all equal. In other words, if the null hypothesis is true, any observed differences in the sample data are merely the result of random chance.

The **alternative hypothesis** (H_a), conversely, suggests that at least one of the groups comes from a different population distribution than the others. It is important to note that the alternative hypothesis does not claim that **all** groups are different from each other; it only claims that **at least one** group is distinct. This distinction is subtle but critical for the correct interpretation of the omnibus test results. If the researcher finds evidence to support the alternative hypothesis, they have discovered a potential effect worth further investigation.

To test these hypotheses, the researcher utilizes the H-statistic calculated from the ranked data. This involves entering the pain ratings into a specialized tool to compute the variance of the ranks. Below is an example of what a typical calculation interface might look like during this process:

This Kruskal-Wallis Test calculator compares the medians of three or more independent samples. It is the nonparametric version of the One-Way ANOVA.

Simply enter the values for up to five samples into the cells below, then press the "Calculate" button.

Sample 1

78, 65, 63, 44, 50, 78, 70, 61, 50, 44

Sample 2

71, 66, 56, 40, 55, 31, 45, 66, 47, 42

Sample 3

57, 88, 58, 78, 65, 61, 62, 44, 48, 77

Once the data is entered, the calculation button is triggered to generate the test statistic and the associated p-value. This step transitions the research from data collection to formal inference, allowing the researcher to move toward a definitive conclusion regarding the efficacy of the different drugs being studied.

CALCULATE

H Statistic: 3.08903

p-value: 0.21342

Interpretive Logic and Drawing Scientific Conclusions

After running the **Kruskal-Wallis test**, the final step is the interpretation of the output. In our knee pain example, the statistical calculation produced a p-value of **0.21342**. To make a decision, the researcher compares this p-value to the pre-established **significance level** (alpha) of 0.05. Because 0.21342 is considerably larger than 0.05, the researcher fails to reject the **null hypothesis**. This is a critical moment in the analysis where the data indicates that the observed differences are not strong enough to be considered statistically significant.

The conclusion drawn from this result is that there is insufficient evidence to suggest that Drug 1, Drug 2, and Drug 3 have different effects on knee pain. While the raw averages of the pain scores might have looked different on the surface, the **Kruskal-Wallis test** revealed that these differences could easily have occurred by chance. This prevents the researcher from mistakenly claiming that one drug is superior to another when the data does not truly support such a claim.

It is also worth noting that failing to find a significant difference does not necessarily mean the drugs are identical; it simply means the current study did not detect a difference. This could be due to a variety of factors, such as a small sample size (low **statistical power**) or very high variability within the groups themselves. In scientific reporting, it is essential to communicate these results clearly, stating that the **medians** were not found to be significantly different under the conditions of the test.

For those interested in performing these analyses themselves, various software packages offer streamlined workflows. Whether using Python, SPSS, or SAS, the underlying logic of the **Kruskal-Wallis test** remains the same. Mastering this test allows researchers to confidently handle complex datasets and contribute valuable, statistically sound insights to their respective fields of study.

Python implementation usually involves the SciPy library.

SPSS provides a user-friendly GUI for non-parametric tests.

SAS is frequently used in clinical and pharmaceutical industries for these calculations.