How to Perform a Kruskal-Wallis Test in Stata: A Step-by-Step Guide

Authored by

stats writer

March 8, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Perform a Kruskal-Wallis Test in Stata: A Step-by-Step Guide*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=134669

The **Kruskal-Wallis Test** is a powerful **non-parametric** statistical procedure used to determine if there are statistically significant differences between the **medians** of three or more independent groups. Unlike the standard **One-Way ANOVA**, which relies on the assumption of **normality**, the Kruskal-Wallis Test does not require the data to follow a specific distribution. This makes it an essential tool for researchers working with **ordinal data** or continuous data that exhibit significant **skewness** or **outliers**. By transforming raw scores into **ranks**, this test evaluates whether the samples originate from the same distribution, providing a robust alternative when parametric assumptions are violated.

In Stata, performing this analysis is highly efficient due to the built-in **kwallis** command. This command calculates the necessary **rank sums** and provides a **test statistic** that follows a **chi-squared distribution**. Using **Stata** for this purpose allows for a seamless workflow that transitions from **data cleaning** to **descriptive statistics** and finally to **inferential testing**. Because the Kruskal-Wallis Test is a **rank-based** method, it is particularly resilient to the influence of extreme values, ensuring that your **statistical conclusions** remain valid even in the presence of irregular data points.

The primary **null hypothesis** for this test posits that the **population medians** of all groups are identical. Conversely, the **alternative hypothesis** suggests that at least one group median differs significantly from the others. Throughout this tutorial, we will explore the nuances of executing this test within the **Stata** environment, interpreting the resulting **p-values**, and reporting the findings with academic precision. By following these steps, you will gain a deeper understanding of how to manage **multi-group comparisons** in a **non-parametric** framework.

Theoretical Foundations of the Kruskal-Wallis Test

Before diving into the software implementation, it is crucial to understand the mathematical logic behind the Kruskal-Wallis Test. This test is essentially an extension of the **Mann-Whitney U Test** for more than two groups. It works by pooling all observations from every group and ranking them from smallest to largest. If the **null hypothesis** is true, the average **ranks** for each group should be approximately equal. If certain groups have consistently higher or lower **ranks**, the resulting **H-statistic** will increase, leading to a smaller p-value.

One of the key requirements for a meaningful Kruskal-Wallis Test is the assumption of **independent observations**. Each subject or data point must belong to only one group, and there should be no relationship between the observations in different groups. Additionally, while the test does not assume **normality**, it does assume that the distributions of the groups have a similar shape. If the shapes are identical, the test specifically compares **medians**; if the shapes differ, the test compares the **mean ranks** of the distributions.

Researchers often turn to **non-parametric** methods like the Kruskal-Wallis Test when dealing with

Likert scales or small sample sizes where **Central Limit Theorem** effects may not yet be apparent. It serves as a safeguard against **Type I errors** that can occur when **ANOVA** is applied to heavily skewed data. By focusing on the order of the data rather than the precise intervals between values, the test remains highly reliable in diverse **experimental designs** and **observational studies**.

Setting Up the Stata Workspace and Loading Data

To begin our practical demonstration, we must first initialize the **Stata** environment and load a suitable **dataset**. For this tutorial, we will utilize the **census** dataset provided by **Stata-Press**, which contains **demographic** information from the 1980 United States **census**. This dataset is ideal because it includes **categorical variables** (such as geographic regions) and **continuous variables** (such as median age), which are necessary for performing a **group comparison**.

Load the dataset by entering the following command into your **Stata Command box**:

use http://www.stata-press.com/data/r13/census

Once the data is loaded, it is standard practice to perform an initial **data audit**. This involves checking for missing values and ensuring that the **variable types** are correctly assigned. The **census** data categorizes states into four distinct **regions**: the Northeast, North Central, South, and West. Our objective is to determine if the **median age** of the population varies significantly across these different geographic areas. Organizing your **Stata** workspace effectively at this stage prevents errors during the **statistical computation** phase.

After loading the data, you should generate a **statistical summary** to understand the characteristics of your **variables**. This provides a baseline for the **descriptive statistics** that will eventually accompany your **inferential test** results. Use the **summarize** command to view the **mean**, **standard deviation**, and range of the variables in the dataset:

summarize

```
. use http://www.stata-press.com/data/r13/census
(1980 Census data by state)

. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       state |          0
      state2 |          0
      region |         50        2.66    1.061574          1          4
         pop |         50     4518149     4715038     401851   2.37e+07
      poplt5 |         50    326277.8    331585.1      35998    1708400
-------------+---------------------------------------------------------
     pop5_17 |         50    945951.6    959372.8      91796    4680558
      pop18p |         50     3245920     3430531     271106   1.73e+07
      pop65p |         50    509502.8    538932.4      11547    2414250
    popurban |         50     3328253     4090178     172735   2.16e+07
      medage |         50       29.54    1.693445       24.2       34.7
-------------+---------------------------------------------------------
       death |         50    39474.26    41742.35       1604     186428
    marriage |         50     47701.4    45130.42       4437     210864
     divorce |         50    23679.44    25094.01       2142     133541
```

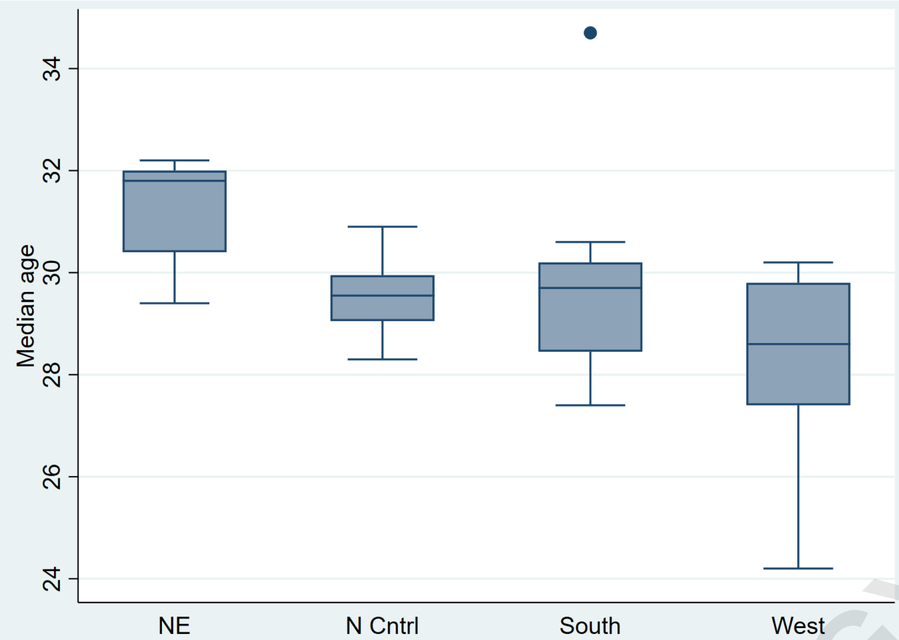
In the resulting output, identify the **medage** (median age) and **region** variables. These are the primary components of our analysis. The **medage** variable represents our **dependent variable** (measurement), while the **region** variable serves as our **independent variable** (grouping factor).

Exploratory Data Analysis and Visualization

Before conducting the formal Kruskal-Wallis Test, it is highly recommended to **visualize** the data. **Exploratory Data Analysis (EDA)** allows you to identify **outliers** and observe the **distribution** of the **dependent variable** across different groups. In **Stata**, one of the most effective ways to compare groups visually is through a box plot. A **box plot** displays the **median**, **quartiles**, and potential **outliers**, providing a clear picture of the **variance** within each category.

To create a **box plot** of median age categorized by region, execute the following command:

graph box medage, over(region)



By examining the resulting **graph**, you can observe whether the **interquartile ranges** overlap significantly. If the **median** lines within the boxes are at different levels, it suggests that a **statistically significant difference** may exist. This visual evidence serves as a preliminary check and helps justify the use of a **non-parametric** test if the distributions appear **non-normal** or if there is **heteroscedasticity** (unequal variances) between the regions.

Visualizing your data also helps in detecting **data entry errors** or extreme **outliers** that could disproportionately affect **parametric tests**. Since the Kruskal-Wallis Test relies on **ranks**, it is less sensitive to these **outliers**, but knowing they exist is vital for a comprehensive **data analysis**. **Stata**'s graphing capabilities are robust, allowing you to customize labels and titles to make the **visualization** ready for **publication** or presentation.

Executing the Kruskal-Wallis Command in Stata

Once you are satisfied with your **exploratory analysis**, you can proceed to the core **statistical test**. The **syntax** for the Kruskal-Wallis Test in **Stata** is straightforward. You use the **kwallis** command, followed by the **measurement variable** and then a comma with the **by()** option to specify the **grouping variable**. This structure tells **Stata** exactly which **metric** to rank and which **categories** to compare.

For our specific **census** example, the syntax is as follows:

kwallis medage, by(region)

```
Kruskal-Wallis equality-of-populations rank test
```

region	Obs	Rank Sum
NE	9	376.50
N Cntrl	12	294.00
South	16	398.00
West	13	206.50

```
chi-squared =     17.041 with 3 d.f.
probability =      0.0007

chi-squared with ties =     17.062 with 3 d.f.
probability =      0.0007
```

When you run this command, **Stata** processes the **rank sums** for each **region**. It calculates the **expected rank sum** under the **null hypothesis** and compares it to the **observed rank sum**. The **algorithm** automatically handles the assignment of **ranks**, including the treatment of **tied observations**, which is a critical step in ensuring the **accuracy** of the **H-statistic**. This automation reduces the risk of manual **calculation errors** that often occur when performing **rank-based tests** by hand.

The output generated by **Stata** provides a comprehensive **summary table**. This table lists each group, the number of **observations** (n) within that group, and the **rank sum**. These details are essential for verifying that the **test** has been applied to the correct **sample size** and for understanding which groups contributed most to the overall **test statistic**. Following the execution of the command, the next vital phase is the **interpretation** of these **numerical results**.

Interpreting the Statistical Output

Interpreting the Kruskal-Wallis Test output requires a focus on three primary values: the **rank sums**, the **Chi-squared statistic**, and the **probability value**. The **rank sums** indicate the total value of **ranks** assigned to each region. A higher **rank sum** relative to the **sample size** suggests that the group generally contains higher values for the **measurement variable**. In our example, **Stata** displays these totals clearly, allowing for an immediate **qualitative comparison** of the regions.

The **Chi-squared with ties** value is our **test statistic** (H). In this analysis, the value is **17.062**. This number represents the magnitude of the difference between the **observed ranks** and what would be expected if all **group medians** were identical. A larger **Chi-squared** value indicates a greater discrepancy between the groups. **Stata** also provides the **degrees of freedom** (df), which is calculated as the number of groups minus one (k - 1). With four regions, our **degrees of freedom** is 3.

The most critical component for **hypothesis testing** is the **probability (p-value)**. In our **Stata** output, the **p-value** is **0.0007**. In **statistical inference**, we typically compare this value to a **significance level** (alpha), usually set at **0.05**. Since **0.0007** is significantly lower than **0.05**, we have sufficient **empirical evidence** to **reject the null hypothesis**. This leads us to conclude that there is a **statistically significant difference** in the **median age** across at least two of the four **census regions**.

Post-hoc Testing and Further Analysis

While the Kruskal-Wallis Test tells us that a difference exists, it does not specify *which* groups are different from one another. This is known as an **omnibus test**. To pinpoint the exact **pairwise differences**, researchers must perform **post-hoc analysis**. Common choices for **post-hoc** testing following a **non-parametric** omnibus test include **Dunn's Test** or multiple **Mann-Whitney U tests** with a **Bonferroni correction** to control for **Type I error inflation**.

In **Stata**, you can install user-written commands like **dunntest** to perform these comparisons. These tests evaluate each possible pair of groups (e.g., Northeast vs. South, West vs. North Central) to see where the **statistical significance** lies. Without **post-hoc testing**, your analysis remains incomplete, as you can only state that the regions are "not all the same" rather than identifying the specific **geographic trends** in **median age**.

Additionally, it is important to consider the **effect size**. While the **p-value** tells you if a difference is likely due to chance, the **effect size** (such as **epsilon-squared**) tells you the **magnitude** of that difference. Reporting **effect sizes** alongside **p-values** is increasingly required by **academic journals** to provide a more nuanced view of the **practical significance** of the research findings. **Stata** provides the foundational data needed to calculate these metrics manually or through additional **plugins**.

Reporting Your Findings Formaly

The final step in the **research process** is to report the results in a clear, professional manner suitable for a **manuscript** or **technical report**. A standard **statistical report** should include the name of the test performed, the **test statistic**, the **degrees of freedom**, the **p-value**, and a clear statement regarding the **null hypothesis**. It is also helpful to include the **sample sizes** (n) for each group to provide context for the **rank sums**.

Based on our **Stata** analysis, an appropriate write-up would be: "A **Kruskal-Wallis Test** was conducted to examine whether **median age** differed significantly across four United States regions: the Northeast (n = 9), North Central (n = 12), South (n = 16), and West (n = 13). The results indicated a **statistically significant difference** in median age between the regions (χ^2**(3) = 17.062**, **p = 0.0007**). Consequently, the **null hypothesis** of equal medians was rejected, suggesting that **geographic location** has a significant impact on the age distribution within the 1980 census data."

When **reporting results**, always ensure that your **notations** follow standard **APA** or **Vancouver** guidelines. Using the **Greek letter** for **Chi-squared** (χ^2) and clearly stating your **alpha level** adds to the **credibility** of your work. By combining **descriptive statistics** (from Step 1), **visualizations** (from Step 2), and **inferential results** (from Step 3), you create a compelling and **rigorous narrative** of your **data analysis** journey in **Stata**.

Best Practices for Non-Parametric Analysis

To ensure the **integrity** of your **statistical analysis**, always verify that your **data structure** is appropriate for the Kruskal-Wallis Test. Ensure that your **dependent variable** is at least **ordinal** and that your **groups** are truly **independent**. If your groups are **related** or **matched** (such as in a **pre-test/post-test** design), you should use the **Friedman Test** instead of the Kruskal-Wallis Test.

Moreover, be mindful of **sample size**. While **non-parametric tests** are excellent for small samples, they generally have less **statistical power** than **parametric tests** when the assumptions of **normality** are met. If your data is **normally distributed**, an **ANOVA** might be more likely to detect a **true effect**. Therefore, always perform **normality tests** (like the **Shapiro-Wilk test**) before deciding which path to take in your **statistical workflow**.

Finally, utilize **Stata**'s extensive **documentation** and help files. By typing **help kwallis** in the **Command box**, you can access detailed information about the **mathematical formulas** used and additional options available for the command. Staying informed about the **latest updates** in **statistical software** ensures that your **methodology** remains **current** and **defensible** in the **scientific community**. With these tools and techniques, you are well-equipped to perform **complex group comparisons** with confidence.