How to Easily Compare Groups with the Kruskal-Wallis Test in Stata

Authored by

stats writer

December 28, 2025

The Kruskal-Wallis Test is a fundamental non-parametric test widely utilized in statistics. Its primary function is to determine if there are statistically significant differences among the medians of two or more independent groups. This test is essential when the assumptions required for standard parametric tests, such as one-way ANOVA, cannot be met--specifically, when data is not normally distributed or when the measurements are only on an ordinal scale.

In the Stata statistical software package, the Kruskal-Wallis Test is executed using the succinct `kwallis` command. This powerful command processes the input data and generates the crucial test statistic, the corresponding p-value, and a comprehensive summary of the results. It is important for analysts to understand that while the Kruskal-Wallis test is flexible, it still requires the data to consist of independent observations across groups. Furthermore, the dependent variable must be continuous or, at minimum, measured on an ordinal scale, allowing for meaningful ranking, which is the operational basis of this particular test.

Understanding the appropriate context for the Kruskal-Wallis Test is key to proper statistical analysis. It serves as the direct non-parametric alternative to the one-way ANOVA. When you have three or more groups, and you want to compare their central tendencies without assuming normality or homogeneity of variances, the Kruskal-Wallis Test provides a robust and reliable methodology.

A Kruskal-Wallis Test is utilized to determine whether or not there is a statistically significant difference between the medians of three or more independent groups. It is considered to be the non-parametric equivalent of the One-Way ANOVA.

This tutorial provides a detailed walkthrough explaining how to conduct a Kruskal-Wallis Test in Stata, focusing on the steps from data loading and visualization to execution and result interpretation.

Selecting the Appropriate Statistical Test

Before diving into the execution, it is critical to confirm that the Kruskal-Wallis Test is the most appropriate statistical procedure for your data. This test operates on the principle of ranking observations across all groups combined, comparing the mean ranks of the groups rather than the means themselves. This ranking process makes it resilient to outliers and deviations from the normal distribution, which would otherwise violate the assumptions of parametric tests.

The key assumptions for the Kruskal-Wallis Test are straightforward: the variable of interest should be measured at an ordinal, interval, or ratio level (but often applied when interval/ratio data is non-normal); the groups must be independent; and the distributions of all groups must have the same shape, even if they are not normal. If the distributions have different shapes, then the Kruskal-Wallis Test still indicates differences, but these differences might not solely be attributed to the

medians, requiring more cautious interpretation.

By choosing this test, we are essentially testing the null hypothesis that the population medians (or, more precisely, the mean ranks) are equal across all comparison groups. If the resulting p-value is sufficiently small (typically less than 0.05), we reject the null hypothesis, concluding that at least one group median differs from the others.

Case Study: Median Age Across US Regions

For this practical example, we will employ the well-known *census* dataset, which is readily available within the Stata documentation and contains 1980 census data for all fifty states in the U.S. This dataset allows us to explore regional demographic differences. Specifically, the states are categorized into four distinct geographical regions, serving as our independent grouping variable:

Northeast
North Central
South
West

The dependent variable we will analyze is *medage*, which represents the median age of the population in each state. Our primary objective is to use the Kruskal-Wallis Test to statistically determine whether the population median age is equal across these four major regions of the United States. This application is perfect for the test, as we are comparing three or more independent groups using a continuous variable (median age).

Analyzing demographic data often involves variables that may not perfectly adhere to the normal distribution, especially when dealing with smaller samples within each group (like the number of states per region). This potential non-normality justifies the use of a non-parametric method like the Kruskal-Wallis Test over its parametric counterpart, ANOVA, ensuring the validity of our conclusions regarding regional differences in central tendency.

Step 1: Loading and Summarizing the Data in Stata

The initial step in any statistical analysis within Stata is loading the necessary data. Since the *census* dataset is a standard example file hosted by StataPress, it can be loaded directly from the web using the `use` command, which saves time and effort compared to manually importing a file. Execute the following command in the Stata Command window to retrieve the dataset:

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

Once the data is successfully loaded, it is good practice to get a rapid overview of the dataset's

structure, variable types, and missing values. This step confirms the data is loaded correctly and identifies the key variables needed for the analysis. We can obtain a quick summary of all variables in the dataset by using the following descriptive command:

summarize

The output generated by the `summarize` command provides essential descriptive statistics, including the number of observations, mean, standard deviation, minimum, and maximum values for all numerical variables. Reviewing this output helps confirm the sample size (N=50 states) and the range of our primary dependent variable, *medage*.

```
. 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

Based on the summarized output, we can observe that the dataset contains thirteen different variables. For the purpose of the Kruskal-Wallis Test, we are only interested in two specific variables: `medage` (representing the median age of the state population, our measurement variable) and `region` (the grouping variable classifying the states into four distinct geographical regions). We must ensure that `region` is coded correctly as a categorical variable for the test to function properly.

Step 2: Visualizing Group Distributions Using Box Plots

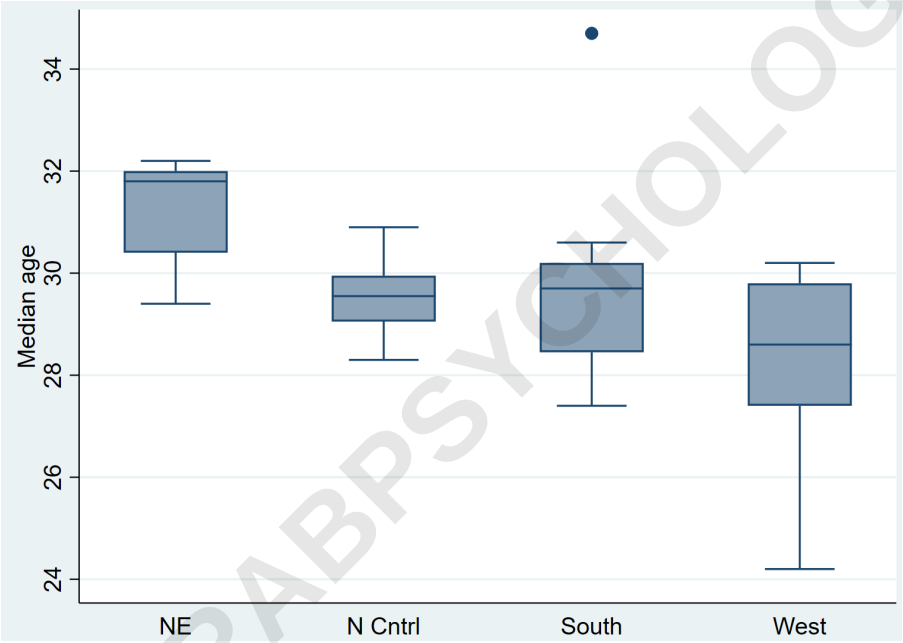
Prior to executing any inferential statistical test, robust data visualization is essential. Visualizing the data provides intuitive insight into the central tendency, spread, and shape of the distributions for each group, which is particularly important when applying a non-parametric test like Kruskal-Wallis. Stata allows for the easy generation of multiple box plots simultaneously, grouped by a

categorical variable.

We will create box plots to visualize the distribution of median age (`medage`) across each of the four regions (`region`). The box plot clearly displays the median (the line inside the box), the interquartile range (the box boundaries), and potential outliers, allowing us to visually assess if the central locations appear different before performing the formal hypothesis test. Execute the following command:

graph box medage, over(region)

Analyzing the resulting graph provides immediate visual evidence regarding the research question. If the median lines in the box plots are vertically misaligned across the regions, it suggests that there might be a statistically significant difference in median age. Conversely, if they are roughly aligned, the null hypothesis of equal medians is more likely to be retained.



Observing the box plots above, there appears to be a noticeable variation in the central location of median age among the four regions, particularly the Northeast region which seems to exhibit a higher median age compared to the South and West regions. This visual evidence strengthens the need for the formal Kruskal-Wallis Test to quantify this difference.

Step 3: Executing the Kruskal-Wallis Test in Stata

The execution of the Kruskal-Wallis Test in Stata is performed using the `kwallis` command. This command requires the specification of the measurement variable (the dependent variable) followed by the `by()` option, which contains the grouping variable (the independent variable). The general

syntax is intuitive and straightforward:

kwallis measurement_variable, by(grouping_variable)

In the context of our census data example, where we are testing the median age (`medage`) grouped by region (`region`), the specific command syntax to be entered into Stata is as follows:

kwallis medage, by(region)

Upon execution, Stata calculates the rank sums for each group, the overall test statistic (H statistic, approximated by Chi-squared), the degrees of freedom, and the critical p-value required for hypothesis testing. It is crucial to ensure that the grouping variable specified in the `by()` option is truly categorical and represents the independent groups being compared.

```
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
```

The resulting output provides all the statistical data necessary to make an informed decision regarding the null hypothesis of no difference in median age across the four regions. The next step involves dissecting this output to understand what the calculated values mean in practical terms.

Step 4: Interpreting the Stata Output

The output of the `kwallis` command is structured to facilitate clear interpretation. It typically presents two main components: a summary table detailing the groups, and the inferential statistics at the bottom. Here is how to interpret the key figures:

Summary table: This initial table provides a breakdown of the sample size (N, or "Obs") for each region (Northeast, North Central, South, West) and, most importantly, the average rank sum for

each region. The Kruskal-Wallis Test operates entirely on these ranks. If the null hypothesis were true, we would expect the mean rank sums for all regions to be approximately equal. Differences in these rank sums indicate potential disparities in the location of the median age distribution across the groups.

Chi-squared with ties: This value represents the calculated test statistic (H). In this case, the Chi-squared value is 17.062. This statistic measures the extent to which the differences observed in the rank sums deviate from what would be expected if the null hypothesis were true. A higher Chi-squared value suggests greater differences among the groups. Stata adjusts this calculation for ties in the data, ensuring a more accurate result. The corresponding degrees of freedom (df) is equal to the number of groups minus one (4 - 1 = 3).

Probability: This is the p-value associated with the calculated Chi-squared test statistic. For our analysis, the p-value turns out to be 0.0007. This value represents the probability of observing our data (or data more extreme) if the null hypothesis--that the median age is equal across all four regions--were true. Since 0.0007 is significantly less than the conventional alpha level of 0.05, we possess strong statistical evidence to reject the null hypothesis. We must conclude that the population median age is not equal across all four regions; there is a statistically significant difference in median age among at least two of the regions.

Step 5: Reporting the Statistical Results

The final and equally important step is the professional reporting of the results. When presenting findings from a Kruskal-Wallis Test, standard reporting practices require including the descriptive statistics for each group (such as sample size), the test statistic (Chi-squared), the degrees of freedom (df), and the exact p-value. This ensures transparency and replicability.

Since the Kruskal-Wallis test is an omnibus test, a significant result only tells us that differences exist somewhere among the groups; it does not specify which particular pairs of regions differ from each other. If the test is significant, a researcher typically follows up with post-hoc pairwise comparisons (such as Dunn's test, which may require additional commands or user-written packages in Stata) to identify the specific group differences.

Below is an example of how to formally report the findings from our analysis of the census data, structured to meet academic and professional standards:

A Kruskal-Wallis Test was performed to determine if the median age of individuals was the same across the following four independent regions in the United States:

Northeast (n = 9)
North Central (n = 12)

South (n = 16)
West (n = 13)

The test revealed a statistically significant difference in the distribution of median age across the four regions, thereby rejecting the null hypothesis of equal population medians ($X2(3) = 17.062$, $p = 0.0007$). Specifically, there was a statistically significant difference in median age between two or more of the regions. Further post-hoc analysis would be required to identify which specific regional pairs differ significantly.