

A Step-by-Step Guide to the Kruskal-Wallis Test in Stata

Authored by
Mohammed loot

November 8, 2025

RECOMMENDED CITATION

Mohammed loot (2025). *A Step-by-Step Guide to the Kruskal-Wallis Test in Stata*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13648>

The [Kruskal-Wallis Test](#) stands as a cornerstone in statistical methodology, essential for determining whether statistically significant differences exist among the medians of three or more independent groups. Its utility stems from its role as the direct [non-parametric](#) alternative to the standard [one-way analysis of variance \(ANOVA\)](#), making it invaluable in situations where parametric assumptions are violated.

Crucially, the Kruskal-Wallis test bypasses the stringent requirement that data must be normally distributed, a necessity for ANOVA. This flexibility makes it indispensable for analyzing data that is severely skewed, contains significant outliers, or consists of inherently ordinal measurements. This comprehensive guide provides an expert, step-by-step methodology for correctly executing and interpreting the Kruskal-Wallis Test utilizing the powerful statistical software package, [Stata](#).

Theoretical Foundation: Understanding the Kruskal-Wallis H Test

The Kruskal-Wallis H test operates on the principle of comparing the mean ranks assigned to observations across the different groups, rather than comparing the means or medians directly. When the test yields a statistically significant result, it provides strong evidence that at least one group's distribution stochastically dominates another group's distribution. Conceptually, we are testing the critical research question: Are the population medians of all comparison groups equal? This forms the basis of the [null hypothesis](#), which we test against the alternative hypothesis that at least one population median differs significantly from the others.

Choosing the Kruskal-Wallis test is appropriate when your research involves analyzing a continuous or ordinal dependent variable categorized across multiple independent groups, particularly when preliminary data analysis suggests that the underlying distributions are highly skewed or fail to meet the assumption of homogeneity of variances required for parametric tests like ANOVA. The robust nature of this non-parametric approach, allowing for the analysis of complex data structures without demanding complex transformations, cements its importance in rigorous statistical research.

It is paramount to recognize the limitations of this test: while it confidently identifies that a difference exists among the groups, it does not specify which specific pairs of groups are the source of that difference. Therefore, if the overall test result is significant, researchers must typically follow up with dedicated post-hoc comparisons--such as Dunn's test or similar multiple comparison procedures, often implemented through additional commands in statistical packages--to pinpoint the exact location of the statistical difference.

Setting Up the Case Study in Stata

To demonstrate the practical application and execution of the Kruskal-Wallis test, we will employ a publicly available dataset bundled with Stata: the standard **census** data. This dataset

encapsulates vital demographic information related to the 1980 U.S. census for all fifty states. Our specific objective is to determine if the median age of residents varies significantly across the four major geographical regions defined within this dataset. The critical variables for our analysis are the independent grouping variable, **region**, and the dependent measurement variable, **medage** (median age).

The states within this dataset are meticulously categorized into four mutually exclusive and exhaustive geographic designations:

Northeast

North Central

South

West

Our analysis hinges on testing the hypothesis that the median age of the population is statistically equivalent across these four geographic areas. Successfully conducting this comparison is essential for understanding and quantifying regional demographic disparities that existed during that census period.

Step 1: Load and Inspect the Data.

The foundational step requires seamlessly loading the necessary dataset directly into the active Stata environment. To retrieve the data directly from the Stata Press repository, execute the following command precisely in the Command box:

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

Following data loading, it is considered best practice in statistical analysis to obtain a preliminary summary of the variables. This step ensures data integrity, confirms variable types, and provides an immediate understanding of the structure and range of the variables. Use the simple but effective `summarize` command for this essential overview:

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

The resulting summary output confirms the presence and basic statistics of all variables in the dataset. For the dedicated purpose of this Kruskal-Wallis analysis, we will exclusively focus on two specific variables: *medage*, which represents the median age of each state's population, and *region*, which serves as our categorical grouping variable.

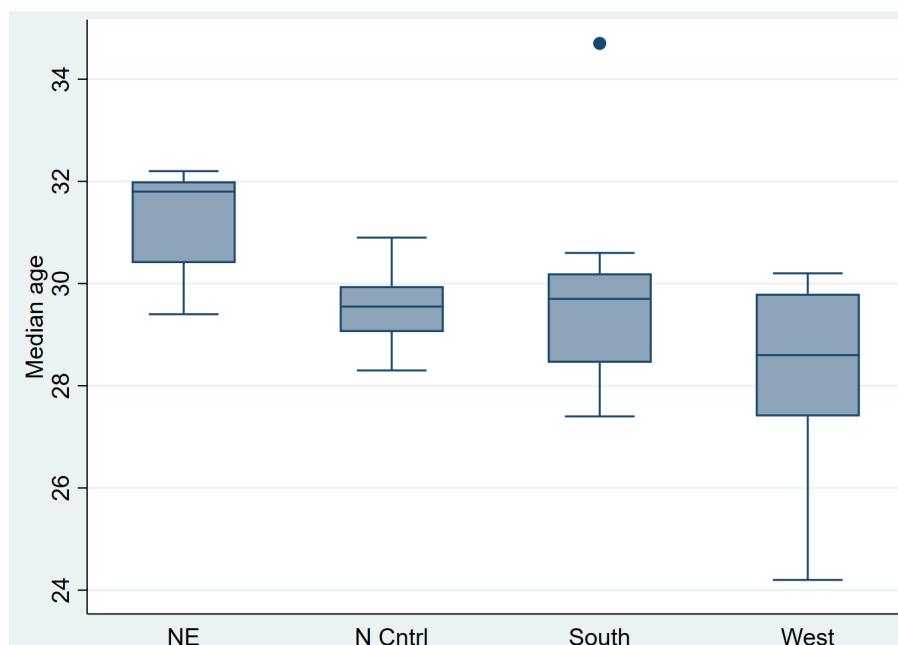
Exploratory Data Analysis: Visualizing Distributions Across Groups

Before proceeding with any formal inferential statistical testing, particularly non-parametric tests which are sensitive to distribution shape, it is critically important to visualize the raw data distribution. This exploratory data analysis (EDA) step offers immediate, intuitive insight into the spread, central tendency, and potential existence of significant outliers within each group. Visualization reinforces the decision to utilize a non-parametric approach and helps anticipate the test outcome.

Step 2: Visualize the Data.

To effectively visualize the distribution of median age across the four geographical regions, we will generate comparative [boxplots](#). Boxplots are exceptionally well-suited for comparing distributions because they clearly delineate the median (the dividing line inside the box), the interquartile range, and the overall spread and potential outliers for each group simultaneously. Execute the following standard Stata command:

```
graph box medage, over(region)
```



Inspection of the resulting graph, which displays four distinct boxplots, reveals tangible differences in both the central tendency (median) and the overall variability of the median age across the regions. For example, a quick visual comparison shows that the Southern region appears to have a visibly higher median age compared to the Western region. This preliminary finding strongly suggests that meaningful regional disparities in age distribution likely exist, providing a strong rationale for the subsequent formal statistical test.

Executing the Kruskal-Wallis Test in Stata

Once the essential steps of data loading, inspection, and visualization are complete, performing the core Kruskal-Wallis test itself is highly efficient within Stata, requiring only the dedicated `kwallis` command. This command is specifically designed to calculate the Kruskal-Wallis H statistic based on the assigned ranks of the observations across all specified groups.

Step 3: Perform a Kruskal-Wallis Test.

The general and universal syntax for executing the Kruskal-Wallis Test in Stata adheres to the following structure:

`kwallis measurement_variable, by(grouping_variable)`

Applying this standard syntax to our specific census dataset variables, where *medage* represents the continuous measurement variable and *region* is the categorical grouping variable, we input the following command into the Stata prompt:

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

The resulting output immediately provides all the necessary statistics required for comprehensive interpretation, including a summary table of ranks, the test statistic, and the associated probability value. A thorough understanding of each element within this output is essential for drawing statistically valid conclusions regarding the null hypothesis.

Interpreting the Statistical Results

The output generated by Stata is thoughtfully structured to facilitate the direct and efficient interpretation of the non-parametric test results. The critical focus of the interpretation must center on the calculated test statistic and, most importantly, the corresponding P-value.

The output is logically divided into three key components:

Summary Table: This initial table provides the exact sample size (n) for each region (Northeast: 9, North Central: 12, South: 16, West: 13) and the crucial corresponding **rank sums**. The rank sum is the aggregate total of the ranks assigned to all observations belonging to that specific group. Disparities observed among these rank sums form the mathematical basis for the calculation of the Kruskal-Wallis test statistic.

Chi-squared with ties: This numerical value represents the calculated [Chi-squared](#) test statistic (H), which is an estimate that appropriately accounts for any ties present within the ranked data. In our specific analysis, the resulting test statistic is **17.062**. This value effectively quantifies the magnitude of the observed differences in mean ranks across the four regions.

Probability: This figure is the calculated [p-value](#) corresponding to the Chi-squared statistic. For

this analysis of regional age disparities, the highly significant probability value is **0.0007**.

To finalize the formal statistical decision, we must compare the calculated P-value to a predetermined level of significance, conventionally set at **alpha (\$alpha\$) = 0.05**. The standard decision rule dictates that if the P-value is less than \$alpha\$, we must reject the null hypothesis (\$H_0\$). Since our calculated P-value of 0.0007 is substantially smaller than the 0.05 threshold, we confidently reject \$H_0\$. This rejection leads to the powerful and actionable conclusion that the population medians of age are not statistically equivalent across all four geographic regions.

Formal Reporting of Findings and Conclusion

The indispensable final step in any comprehensive statistical analysis is the clear, concise, and standardized reporting of the results. This ensures that all relevant statistical metrics are included to fully support the conclusion. The formal reporting must explicitly state the specific non-parametric test performed, the variables utilized, the precise sample sizes for each group, the value of the test statistic, the correct degrees of freedom (which is the number of groups minus one, yielding \$4-1=3\$), and the decisive P-value.

Step 4: Report the Results.

Below is an exemplary standard format for formally reporting the findings derived from the Kruskal-Wallis Test, ensuring absolute clarity regarding the group sizes and the outcome of the hypothesis testing:

A Kruskal-Wallis Test was executed to determine if the median age of individuals was consistent across four major geographic regions in the United States. The sample sizes utilized for each region were:

Northeast (n = 9)

North Central (n = 12)

South (n = 16)

West (n = 13)

The statistical analysis conclusively revealed that the median age of individuals was significantly different across the four regions. The test yielded a Chi-squared statistic of $\chi^2(3) = 17.062$, with a corresponding P-value of $p = 0.0007$. Since this P-value is considerably less than the established significance level ($\alpha = 0.05$), we reject the null hypothesis and conclude that there is a statistically significant difference in median age between at least two or more of the regions defined in the census dataset.

It is important to reiterate that while the overall test confirms the existence of regional differences, further dedicated post-hoc testing (e.g., performing multiple comparisons with appropriate P-value

adjustments) would be absolutely necessary to determine exactly which specific pairs of regions exhibit these significant differences in median age distributions.