Differences in Distribution Shapes and Their Effects on Kruskal–Wallis Test Validity

Joseph Mathews

I. Introduction

In 1952, William H. Kruskal and W. Allen Wallis introduced a new test as a nonparametric alternative to the one-way ANOVA test. By the late 1940s, ANOVA was the standard tool used for comparing multiple groups of data, but it relied heavily on assumptions of normality and equal variances. Although other rank-based tests already existed at the time, like Wilcoxon signed-rank and Mann-Whitney U, these tests could only handle two groups. At the time, there existed no general k-sample rank test analogous to ANOVA. As stated in the abstract of their oft-cited article, a hypothesis test to check whether samples are from the same population can be made by "ranking the observations from 1 to $\sum n_i$ (giving each observation in a group of ties the mean of the ranks tied for), finding the C sums of ranks, and computing a statistic H" (Kruskal & Wallis, 1952).

The resulting Kruskal-Wallis test statistic has many advantages in comparison to other rank-based tests; for one, it reduces to the Mann-Whitney test when $k = 2$, another effect of its generalized logic. The KW test is also asymptotically equivalent to ANOVA under normality, and improves the robustness of rank-based testing, especially when aggregated data in practicality tends to be messy and imperfect. E.L. Lehmann later clarified that Kruskal-Wallis (KW) actually tests equality of distributions, rather than medians as many had believed. When comparing differences in treatment with a population of subjects, "there is an important difference between the comparison of treatments that can be assigned to the subjects at will and attributes that are inseparably attached to the subjects… the hypothesis being tested is that of the absence, respectively, of treatment differences or of an association" (Lehmann & D'Abrera, 2006).

More broadly, the common interpretation of the KW test being a "test of medians" is only valid with the assumption that the group distributions being tested have the same shape and spread. In reality, the null hypothesis of the KW test is that all group distributions are identical, or that the medians come from the same distribution. For this reason the KW test is sensitive to differences in distribution; it can reject the null hypothesis even when group medians are equal if the distributions differ in shape, skew, or tail behavior. The purpose of this project, therefore, is to pinpoint how much differences in distribution shape can affect the reliability of the Kruskal-Wallis test, even when group medians are equal.

II. Methods

I defined a function `simulate_kw` which simulates three groups (A, B, and C) of varying distribution shapes and takes an argument `n`, the sample size for each group. By default the sample size is `n = 500`. Our significance level of `alpha = 0.05` serves to measure the significance of our simulated KW tests and how much they are affected by differences in the distribution shape. Also, note that the test is two-sided by default.

Distribution Groups

I generated three different groups of interest to perform a Kruskal-Wallis test on.

- Group A: Normal distribution
- Group B: Log Normal distribution with right skew
- Group C: heavy-tailed t distribution with 3 degrees of freedom

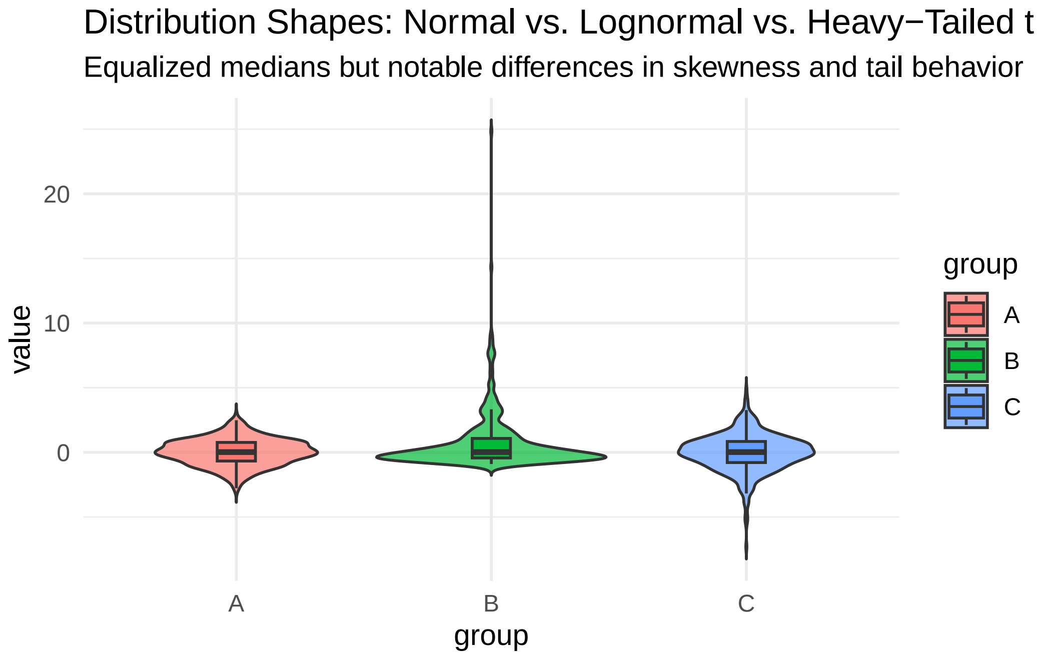
As a baseline, I generated 500 observations for each distribution group (1500 randomly generated observations in total) and ran a KW test on the generated data. As stated previously, the null hypothesis is that all group medians come from the same distribution. If I guarantee that all three group medians are equal, then any rejection of the null hypothesis from the KW test would solely be the result of differences in distribution shape, not just location. I also varied the size of n for each distribution group, which will become relevant later in the paper.

simulate_kw()

First, we simulate `n` observations from a standard Normal distribution, defined as `A`. It serves as the baseline group in the KW test comparisons. We then create `B_raw`, which generates `n` observations from a right-skewed log-normal distribution. We subtract its own median and add the median of group `A` to shift the distribution so that its median matches that of group `A`, creating `B`. This is to ensure that groups only differ in shape, so that any rejection by the Kruskal-Wallis test will be due to the distributions' shape differences rather than any differences in medians. Lastly, we generate `C_raw`, which generates `n` observations from a heavy-tailed t-distribution with 3 degrees of freedom. We perform the same subtracting and adding to create `C` for the same purposes as we did with `B`.

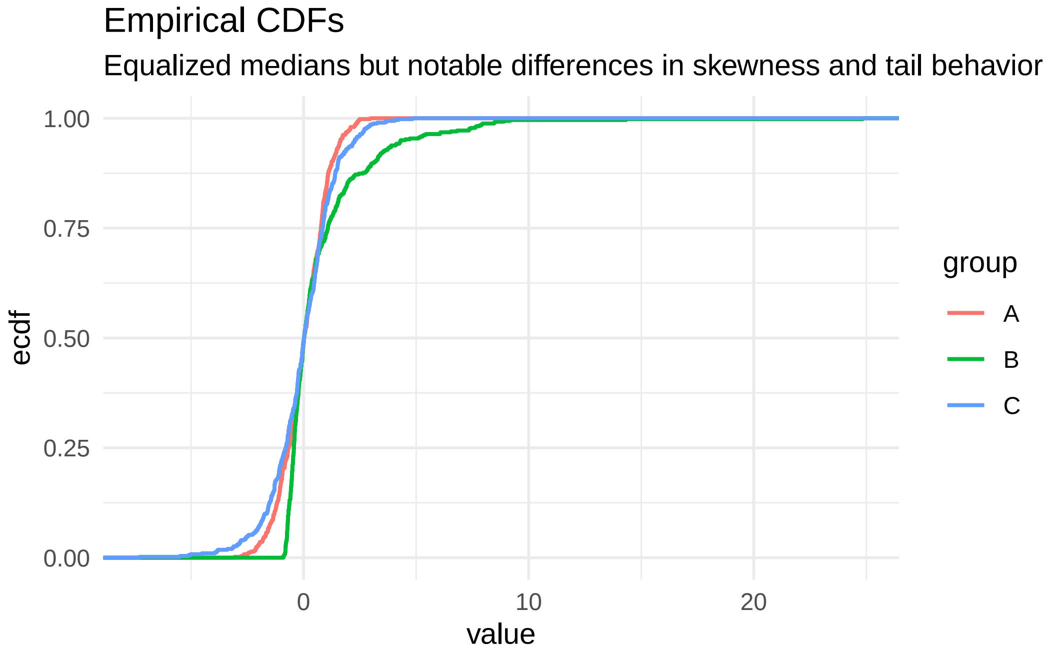
Distribution Plots of A, B, and C

Below is a violin plot of the three groups of interest. Note that although each group has equalized medians, their distribution shapes are quite different. These differences will play an important role in determining the sensitivity of the KW test later on.



Empirical CDFs of A, B, and C

I also created Empirical CDF plots for each distribution group, shown below. Each ECDF intersects at 0.5, indicating that indeed the group medians are equal. However, each ECDF differs greatly in shape: Group A (Normal) and Group C (t-distribution) appear similar but with Group C's heavy-tail behavior drawing the ECDF out a bit further than Group A, whereas Group B (Lognormal) shoots up through the median before slowly growing to 1, a result of its highly skewed distribution.



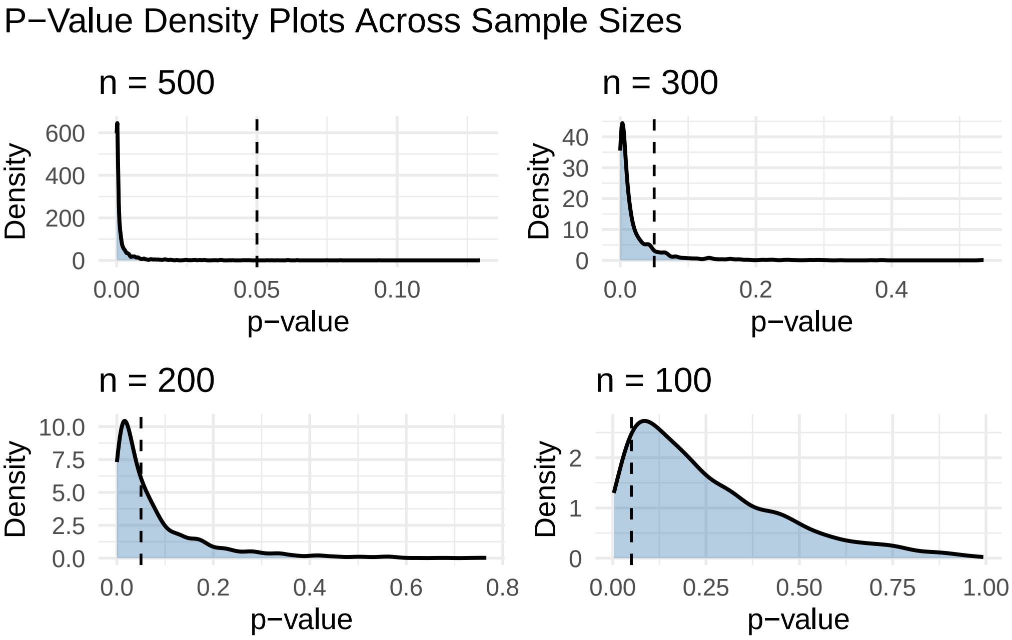
III. Results

Because I am checking how often the Kruskal-Wallis test will falsely reject the null hypothesis due to differences in distribution shape, I created density plots of the resulting p-values from `simulate_kw()`. I also wanted to present the rejection rates in a table to better understand the numerical aspect of the results.

Density Plots

We will vary `n` to see how much the distribution shape affects the Kruskal-Wallis test. We will first use `n = 500` from the newly-created function `simulate_kw()` as a baseline measurement, then vary the sample size to visualize the distribution of the p-values. The dashed line in each plot represents the significance level of $\alpha = 0.05$. Note that anything below the significance level will reject the null hypothesis, even if the medians are equal.

Below are the density plots for the distribution of p-values after running 2000 simulations for each plot.



We can see that the KW will reject the null hypothesis practically every time when $n = 500$. Decreasing the value of n eventually allows for a lessened rejection rate of the null hypothesis, until we reach the fourth plot of $n = 100$, at which point the overwhelming majority of p-values fall above our defined significance level.

Rejection Rates

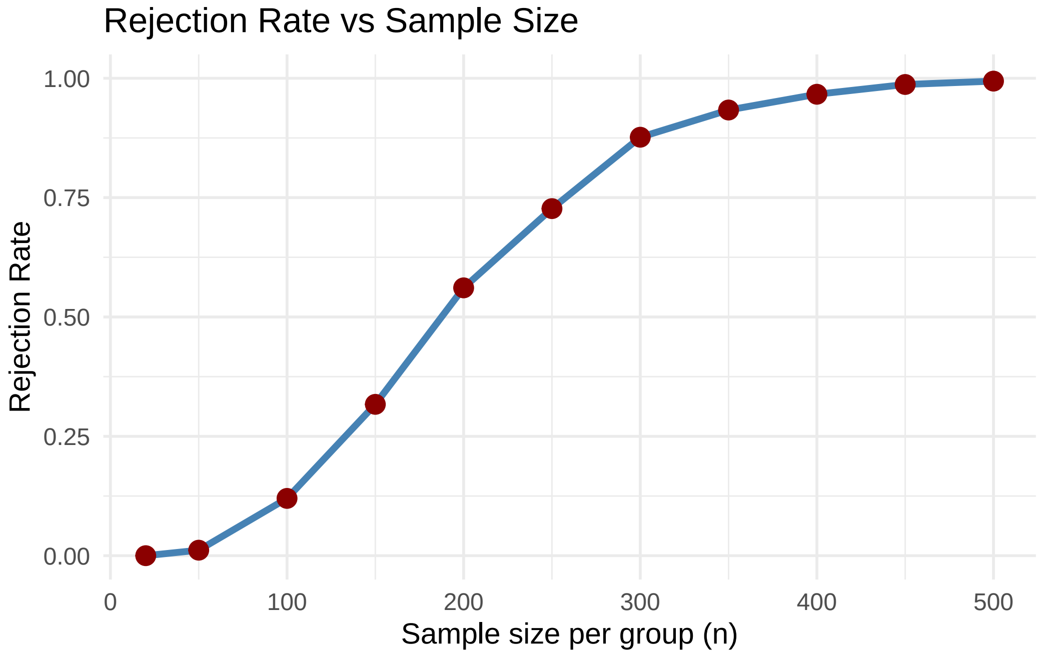
In order to better understand the numerical aspect of our rejection rates, I created a table which displays the rejection rates under equal medians but unequal distributions for varying sizes of n.

Table 1: Rejection Rates vs. Sample Size

n	rejection_rate
20	0.000
50	0.012
100	0.120
150	0.317
200	0.561
250	0.727
300	0.876
350	0.933
400	0.967
450	0.987
500	0.994

The surprising results that, for example, the rejection rate when `n = 500` is 99.4% highlights one of the most fundamental assumptions of the KW test: its null hypothesis assumes equal distributions; more specifically, that all groups come from the same distribution. When that assumption is progressively violated, the KW test, believing that the distribution shapes are equal and finding data which counteracts that underlying assumption, begins to reject the null hypothesis at an increasing rate.

The rejection rates displayed in Table 1 follow an S-shaped power curve when graphed, where the rejection rate increases monotonically with sample size and a sharp increase in rejection rates in the mid-range of sample sizes. At low values of n, the KW test has near-zero power. Rejection rates sharply increase between $n = 100$ and $n = 300$, before plateauing near large values of n and approach 100%.



Skewness and Kurtosis

Analyzing the skewness and kurtosis of each distributional group further emphasizes the difference in shape between each group. In Table 2, Group A's skewness stays close to 0 and its kurtosis stays close to 3 across all sample sizes, serving as a well-behaved reference for the other distributions. Group B consistently shows a large positive skewness due to its right-skewed tail. Larger sample sizes reveal increasingly extreme observations, leading to measured kurtosis values to blow up, becoming especially pronounced with a value of 190.122 for `n = 400`. Group C's skewness tends to stick a little bit below 0, except at `n = 500` where it measures -3.241. Its kurtosis is consistently above 3, highlighting its heavy-tailed nature.

Table 2: Skewness and Kurtosis by Group and Sample Size

group	skewness	kurtosis
n = 50		
A	-0.506	2.416
B	3.390	17.044
C	-0.642	6.069
n = 100		
A	0.014	2.978
B	1.395	5.508
C	-0.492	6.199
n = 200		
A	0.148	3.312
B	3.096	14.536
C	-0.164	7.068
n = 300		
A	0.191	3.095
B	3.247	17.829
C	-0.096	12.531
n = 400		
A	0.035	2.638
B	11.819	190.122
C	0.611	7.377
n = 500		
A	-0.124	3.086
B	3.876	26.849
C	-3.241	41.225

The KW test therefore is highly sensitive to differences in distribution shape, evident in its measured skew and kurtosis for the different distributional groups. Shape differences are incredibly important when considering KW tests and its accompanying interpretations.

IV. Discussion

The analyses presented above clearly show that the Kruskal-Wallis test, although highly advantageous in nonparametric statistical inference, still requires some level of assumptions for proper analysis. Notably, the necessity of the underlying assumption that each group's median must come from the same distribution became increasingly apparent as I increased the sample size for each group. Group A and Group C measured somewhat similarly due to Group C's heavy-tailed distribution still maintaining a semblance to a Normal distribution, but Group B's Lognormal distribution proved to be highly influential in the rejection rates for the KW test, no doubt due to its extreme right-tailed skew. When performing Kruskal-Wallis tests, then, it must be guaranteed that distribution shapes remain as close to equal as possible for effective analyses and useful outputs.

References

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Lehmann, E. L., & D'Abrera, H. J. M. (2006). *Nonparametrics: Statistical methods based on ranks* (Revised 1st). Springer.