Understanding the Friedman Test: A Non-Parametric Approach to Repeated Measures ANOVA in R

Authored by

Mohammed looti

November 8, 2025

The Friedman Test stands as a robust non-parametric alternative to the one-way Repeated Measures ANOVA. This statistical procedure is indispensable when researchers are working with repeated measures designs, meaning the same subjects or matched blocks are evaluated under three or more distinct treatment conditions. The primary goal of the test is to rigorously determine whether statistically significant differences exist among these groups when the underlying assumptions required for traditional parametric tests, particularly the assumption of normality of the differences, cannot be met or are clearly violated.

A key difference between the Friedman Test and its parametric counterpart is its reliance on ranks rather than the raw numerical data. This methodological choice grants the test significant resilience against common statistical challenges, such as the presence of severe outliers or non-normal data distributions, making it exceptionally valuable across diverse fields, including behavioral science, clinical trials, and consumer research. This tutorial aims to provide a comprehensive, step-by-step methodology for correctly structuring, executing, and interpreting this crucial statistical analysis using the powerful R programming environment.

Understanding the Friedman Test and its Application

The fundamental mission of the **Friedman Test** is to investigate whether the underlying distributions, often focusing on the medians, of scores across multiple treatment conditions are statistically identical. Since it belongs to the class of non-parametric statistics, the test operates without demanding restrictive assumptions about the precise shape or parameters of the population distribution from which the samples were drawn. This characteristic is especially advantageous when studies involve small sample sizes or when the data itself is ordinal, meaning the values can be ranked but the distance between them is not necessarily meaningful.

In the context of a typical repeated measures experimental design, subjects act as their own controls, creating inherent dependency among the measurements collected across the different conditions. The Friedman Test meticulously accounts for this dependency by conceptualizing the observations as being grouped into "blocks," where each subject represents a unique block. Within each of these blocks (subjects), the observed scores are transformed into ranks. The final test statistic is then calculated based on how consistently these ranks vary across the different treatment conditions. If the null hypothesis (H_0)--which posits no difference between treatments--is true, the average rank assigned to each treatment condition should be roughly equivalent across all blocks.

Before proceeding with this analysis, researchers must confirm that their study design aligns with a randomized block structure, where the measurements are consistently taken on the same or matched units across at least three distinct conditions. The data input must allow for meaningful ranking (i.e., be at least ordinal). The primary statistical challenge addressed is testing the null

hypothesis (H_0), which states that there are no systematic differences in the median distributions among the groups being compared.

Implementing the Friedman Test in R

The R statistical environment simplifies the execution of this procedure significantly, providing the built-in function `friedman.test()` within its base packages. Successfully conducting the analysis hinges on understanding the function's syntax and correctly mapping the variables from the experimental design onto its required arguments. This mapping ensures that R correctly identifies the response measures, the treatments being compared, and the subject identifiers that define the blocks.

The function requires three essential components, defining the response variable, the grouping factor, and the blocking factor--the foundational elements of any valid repeated measures analysis:

```
friedman.test(y, groups, blocks)
```

Each of the three mandatory arguments plays a specific role in defining the data structure for the R function:

y: This vector contains the **response values** or the measured outcome scores (e.g., physiological measures, reaction times, test performance scores).
groups: This vector identifies the specific **group or treatment condition** associated with each response value. This is the factor whose levels are being compared in the analysis.
blocks: This vector designates the **blocking variable**, which typically serves as the unique identifier for the subject, participant, or matched unit whose measurements are repeated across all treatment conditions.

The result generated by executing this function includes a calculated Chi-Square test statistic and its corresponding p-value. The standard decision rule applies here: if the calculated p-value falls below the predetermined level of significance (conventionally set at $alpha = 0.05$), the researcher must reject the null hypothesis. Rejecting H_0 provides statistical evidence to conclude that the median distributions of the groups are not all equal.

Setting Up the Example Dataset

To illustrate the practical steps of applying `friedman.test()`, we will construct and analyze a simulated dataset focused on measuring the reaction times of participants under the influence of different pharmacological agents. Consider a clinical trial where five distinct patients (blocks) are sequentially administered four different types of drugs (Drug 1, Drug 2, Drug 3, and Drug 4). A specific reaction time (the score) is recorded after each administration. Because every patient is

exposed to all four conditions, this design perfectly satisfies the requirement for a repeated measures analysis suitable for the **Friedman Test**.

In this scenario, our primary analytical objective is to determine whether the average or median reaction time significantly varies depending on the specific drug administered. Consequently, the score represents the response variable, the drug type acts as the grouping variable, and the unique person ID serves as the blocking variable. It is vital to prepare the data in a "long" format, where each row represents a single observation (score), as this is the requisite structure for the base R function.

The following R code snippet demonstrates the creation of this structured dataset, ensuring the proper alignment of the variables needed for the subsequent statistical test:

```
#create data
data <- data.frame(person = rep(1:5, each=4),
drug = rep(c(1, 2, 3, 4), times=5),
score = c(30, 28, 16, 34, 14, 18, 10, 22, 24, 20,
18, 30, 38, 34, 20, 44, 26, 28, 14, 30))

#view data
data
```

person drug score
1 1 1 30
2 1 2 28
3 1 3 16
4 1 4 34
5 2 1 14
6 2 2 18
7 2 3 10
8 2 4 22
9 3 1 24
10 3 2 20
11 3 3 18
12 3 4 30
13 4 1 38
14 4 2 34
15 4 3 20
16 4 4 44
17 5 1 26
18 5 2 28

19 5 3 14
20 5 4 30

This resulting structure, stored in the dataframe named `data`, clearly illustrates that each patient (identified by `person` 1 through 5) contributes a unique reaction `score` for every one of the four drug conditions (1 through 4). This long format is essential because it allows the Friedman Test function to correctly isolate and rank the scores **within** each person, effectively controlling for high inter-individual variability while assessing the differential impact of the drugs.

Executing and Interpreting the Primary Test

Once the data is correctly prepared and structured in the required long format, the execution of the `friedman.test()` is straightforward. We map the variables established in our experimental design directly into the function: `data$score` is specified as the response variable (y), `data$drug` as the grouping factor (groups), and `data$person` as the blocking variable (blocks). This command initiates the non-parametric rank-sum comparison across the four drug groups, ensuring that the unique characteristics of each patient are appropriately managed as blocks.

The R command and subsequent output are as follows:

#perform Friedman Test
friedman.test(y=data$score, groups=data$drug, blocks=data$person)

Friedman rank sum test

data: data$score, data$drug and data$person
Friedman chi-squared = 13.56, df = 3, p-value = 0.00357

The statistical output provides the necessary metrics to evaluate the overall hypothesis. We observe a calculated **Chi-Squared test statistic is 13.56**, determined with 3 degrees of freedom (df). Most critically, the corresponding p-value is calculated as 0.00357. Given that this p-value is significantly less than the standard significance threshold of $alpha = 0.05$, we possess statistically compelling evidence to decisively reject the null hypothesis.

The rejection of H_0 leads to the crucial conclusion that the median reaction times are **not** identical across all four drug treatments. In simpler terms, the type of drug administered results in statistically meaningful differences in patient response time. However, it is important to remember that the Friedman Test is an omnibus test, much like the overall F-test in ANOVA; it only signals that differences exist somewhere among the groups but does not isolate which specific pairs are causing the effect. To pinpoint the exact location of these differences, a subsequent stage of

specific pairwise comparisons--known as post-hoc tests--is required.

Conducting Post-Hoc Analysis

Following the confirmation of a significant overall effect by the primary Friedman Test, the analytical focus must shift to conducting post-hoc tests. These procedures systematically perform pairwise comparisons across all possible combinations of the treatment groups, generating specific p-values for each comparison. This granularity is necessary to fully understand the effects detected in the omnibus test.

The most widely recognized and statistically appropriate post-hoc procedure for the Friedman Test is the pairwise Wilcoxon rank sum test. Since our analysis involves six distinct comparisons (1 vs 2, 1 vs 3, 1 vs 4, 2 vs 3, 2 vs 4, 3 vs 4), running multiple tests substantially elevates the risk of committing a Type I error (a false positive finding). To maintain the integrity of our conclusions by controlling this inflation of the family-wise error rate, it is mandatory to apply a robust p-value adjustment method.

The R function `pairwise.wilcox.test()` is the tool of choice for this procedure, and it offers flexibility in specifying the adjustment method. For this example, we will employ the Bonferroni correction, a highly conservative yet reliable technique known for strictly controlling the probability of finding at least one false positive. The function requires three critical arguments to correctly execute the analysis:

```
pairwise.wilcox.test(x, g, p.adj)
```

x: The response vector (e.g., `data$score`).
g: The grouping vector (e.g., `data$drug`).
p.adj: Specifies the method for adjusting p-values. Common valid string options include `"holm"`, `"hochberg"`, `"hommel"`, `"bonferroni"`, `"BH"`, `"BY"`, `"fdr"`, and `"none"`.

We execute the post-hoc comparison using the specified Bonferroni adjustment:

#perform post-hoc tests
pairwise.wilcox.test(data$score, data$drug, p.adj = "bonf")
Pairwise comparisons using Wilcoxon rank sum test

data: data$score and data$drug

1 2 3
2 1.000 - -
3 0.449 0.210 -
4 1.000 1.000 0.072

P value adjustment method: bonferroni

Interpreting the Post-Hoc Results

The output of the `pairwise.wilcox.test()` function is presented as a matrix that displays the Bonferroni-adjusted p-value for every possible pairing among the four drug groups. Interpreting this matrix requires comparing each adjusted p-value against the chosen significance level ($alpha$). If the adjusted p-value for a specific pair is less than $alpha$, we conclude that those two drug groups exhibit a statistically significant difference in their median reaction times.

A systematic review of the matrix yields the following adjusted probabilities for the six comparisons:

Drug 1 vs. Drug 2: Adjusted p = 1.000 (No significant difference observed)
Drug 1 vs. Drug 3: Adjusted p = 0.449 (No significant difference observed)
Drug 1 vs. Drug 4: Adjusted p = 1.000 (No significant difference observed)
Drug 2 vs. Drug 3: Adjusted p = 0.210 (No significant difference observed)
Drug 2 vs. Drug 4: Adjusted p = 1.000 (No significant difference observed)
Drug 3 vs. Drug 4: Adjusted p = 0.072

When adhering to the most conventional and conservative significance threshold of $alpha = 0.05$, none of the pairwise comparisons achieve statistical significance. However, if the research design permitted a slightly more exploratory threshold, such as $alpha = 0.10$, the comparison between **Drug 3 and Drug 4 (p-value = 0.072)** would be classified as statistically significant. This finding strongly suggests that the difference between these two specific drugs is the primary driver of the overall significant effect initially detected by the primary **Friedman Test**. The rigorous combination of the omnibus test and subsequent, error-controlled post-hoc analysis ensures that the conclusions drawn from repeated measures data are both robust and accurately localized.