How to Perform a Friedman Test to Compare Related Groups

Authored by

stats writer

January 22, 2026

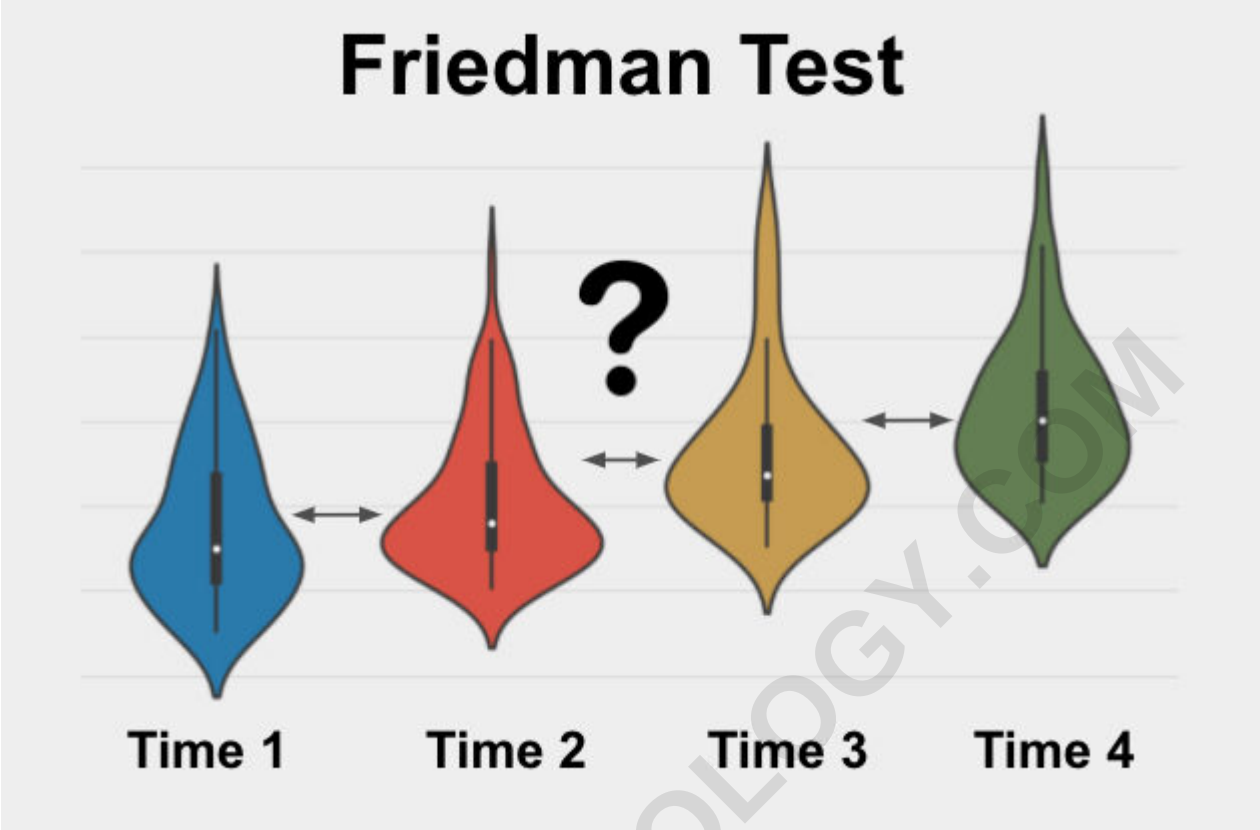
The Friedman Test is a powerful statistical procedure designed to compare the distributions of three or more related samples. Serving as the essential non-parametric alternative to the standard One-Way ANOVA for repeated measures, it is indispensable when analyzing data that violates the stringent assumptions of parametric methods, such as the requirements for normality or homogeneity of variances. The core mechanism of the test involves converting raw scores into ranks within each subject or block, and then analyzing these ranks to ascertain if there is a statistically significant difference across the conditions or time points. This robust method finds wide application across diverse fields, including the social sciences, medicine, psychology, and business research. It was named in honor of its creator, the distinguished American economist and statistician, **Milton Friedman**, who developed the technique in 1937.

Defining the Friedman Test in Modern Statistics

The Friedman Test is fundamentally a hypothesis test used to determine if differences exist among three or more measurements taken from the same group of subjects. This is critical when the dependent variable of interest is measured repeatedly under varying conditions or at multiple time points, making the samples inherently related or dependent. Unlike its parametric counterpart, this procedure does not assume that the data follows a normal distribution, making it highly suitable for variables that are skewed, ordinal, or when dealing with small sample sizes where distributional assumptions cannot be reliably met.

To utilize the Friedman Test effectively, several conditions regarding your data structure must be satisfied. Primarily, your variable of interest must be **continuous** or ordinal, providing meaningful data for ranking. Furthermore, the spread of scores across your related groups should be reasonably comparable, even though the distribution itself does not need to be symmetric. Finally, adequate sample size is important for statistical power; while definitive thresholds vary, having more than five subjects or blocks is generally considered the minimum requirement for producing reliable test results. This focus on ranks allows the test to maintain statistical validity even when the underlying population distribution is unknown or severely non-normal.

The methodology behind the test relies on calculating the rank of each observation relative to the other observations within the same subject or block. By focusing on ranks rather than the raw data values, the test mitigates the influence of extreme outliers or severe violations of parametric assumptions. It then compares the average ranks across the different measurement conditions. If these average ranks differ significantly, it suggests that the experimental manipulation or the passage of time had a genuine effect on the measured variable. The final output is an omnibus test statistic, often approximated by the Chi-square statistic, which indicates whether differences exist somewhere among the groups.



*The Friedman Test is often referred to using alternative terminology that highlights its statistical function. Common aliases include the **Non-Parametric Repeated Measures ANOVA**, the **Non-Parametric Friedman Test**, or simply the **Friedman Rank Sum Test**, all emphasizing its rank-based approach for dependent samples.*

Core Assumptions for the Friedman Test

Before applying any statistical technique, researchers must confirm that their data adheres to the method's underlying assumptions. These prerequisites are not arbitrary; they ensure the validity and accuracy of the resulting statistical inference. Violating these assumptions can lead to misleading conclusions or incorrect interpretations of the differences observed among groups. The Friedman Test, while considered robust due to its non-parametric nature, still relies on specific structural characteristics of the data set that must be satisfied to guarantee that the calculated test statistic accurately follows the known sampling distribution.

Understanding and verifying these requirements is a mandatory step in the data analysis process. The primary assumptions governing the application of the Friedman Test ensure that the ranking procedure is meaningful and that the comparison across conditions is statistically sound. Failure to meet these underlying conditions may necessitate the use of an even more specialized statistical approach or require transforming the data to better meet the test's requirements. These core

assumptions are summarized as follows:

The dependent variable must be measured on at least an **ordinal** scale (often referred to as continuous in statistical software).
The subjects or blocks must represent a **random sample** from the population of interest.
There must be an adequate number of subjects or blocks (**Enough Data**) to achieve sufficient statistical power.

We will now explore each of these crucial assumptions in greater detail to clarify their implications for research design and data collection, ensuring proper methodology implementation.

Requirement 1: Ordinal or Continuous Dependent Variable

The variable being measured (the dependent variable) must be capable of being ordered or ranked meaningfully. Although often referred to simply as a continuous variable in introductory statistics texts, the Friedman Test is truly suitable for data measured on an ordinal, interval, or ratio scale. This flexibility arises because the test operates on ranks. This means the variable can theoretically take on any value within a given range (interval/ratio data), or, at minimum, the differences in magnitude between scores can be consistently ranked (ordinal data). If the variable cannot be ranked (e.g., nominal categories), the test cannot be applied.

Classic examples of suitable data include precise measurements such as age, body weight, height, or standardized test scores. Furthermore, composite scores derived from surveys, particularly those utilizing Likert scales (which are often treated as continuous or interval data in social sciences), are also appropriate, provided the scores allow for meaningful internal ranking within each subject across the repeated measures. The key is that we must be able to assign a relative rank (1st, 2nd, 3rd, etc.) to the measurements taken from the same subject across the different conditions.

*If your dependent variable is confirmed to be approximately normally distributed and meets the assumption of sphericity (the assumption that the variances of the differences between all pairs of related groups are equal), a **One-Way Repeated Measures ANOVA** is the more powerful parametric test you should consider using instead.*

Requirement 2: Simple Random Sample and Related Groups

The integrity of statistical inference hinges on the quality of the sampling procedure. The data points used in the analysis must be derived from a simple random sample of the target population. This ensures that the findings generated from the sample can be validly generalized back to the population from which the sample was drawn. For instance, if a study aims to assess the impact of three different teaching methods on student performance, the group of students selected for the

study must be chosen randomly to be representative of the larger student body.

Furthermore, the fundamental design of the Friedman Test requires the samples to be related or dependent. This dependency arises because the same subjects are measured repeatedly across all conditions. If the selection of subjects was not random, or if there were systematic biases introduced during group assignment or measurement (e.g., selection bias or attrition bias), the subsequent statistical analysis will suffer from **bias**. Bias, in statistical terms, refers to the systematic tendency to obtain incorrect or distorted results due to flaws in the data collection or sampling methodology, compromising the internal and external validity of the study.

If the requirement for a random sample cannot be strictly met, the generalizability of any conclusions must be severely limited and cautiously reported, acknowledging the potential non-representativeness of the sample. Conversely, if you determine that you have three or more measurements derived from different, independent, or unrelated groups, you should employ the ***Kruskal-Wallis One-Way ANOVA****, which is the non-parametric alternative designed specifically for independent samples.*

Requirement 3: Sufficient Sample Size (N > 5)

While the Friedman Test is often favored for smaller datasets compared to parametric tests, a minimum acceptable sample size is necessary to ensure the test statistic accurately approximates the theoretical **Chi-square distribution**, which is used for calculating the p-value. A general consensus among statisticians suggests that the number of subjects or blocks (N) should typically be greater than five in each condition, although some argue for a slightly higher threshold depending on the number of repeated measures (k).

The determination of "enough data" is also intrinsically linked to the concept of **statistical power**--the probability of correctly rejecting a false null hypothesis. If the expected differences (effect size) across the groups are anticipated to be substantial, a smaller sample size may still possess sufficient power to detect that large effect. However, if the expected differences are subtle or small, a significantly larger sample size will be required to detect such minor variations reliably and prevent a Type II error (failing to detect a real effect). Researchers should ideally conduct a power analysis prior to data collection to determine the required sample size based on the anticipated effect size and desired power level.

Applying the Friedman Test: The Decision Framework

The choice of statistical test is determined by the research question, the level of measurement of the variables, and the structure of the data collection design. The Friedman Test provides a unique solution for researchers when they face specific constraints regarding their variables and sampling

methodology. You should confidently select the Friedman Test when your study simultaneously meets the following four key criteria, ensuring both appropriateness and validity of the statistical analysis.

The research objective is to assess whether three or more groups exhibit a significant **difference** in their central tendency.
The outcome measure (dependent variable) is **ordinal** or continuous.
The design includes **three or more groups**, conditions, or time points being compared.
The samples across these groups are unequivocally **related** or dependent (i.e., repeated measures).

A clear understanding of these criteria is essential for selecting the correct non-parametric statistical tool, thereby ensuring that your inferential conclusions are sound and justifiable based on your experimental design. This framework helps differentiate the Friedman Test from alternative procedures like the Wilcoxon Signed-Rank Test or the Kruskal-Wallis Test.

Criterion 1: Testing for Differences in Distributions

The primary purpose of the Friedman Test is comparison. Researchers utilize this test to evaluate the null hypothesis that the population distributions from which the related samples are drawn are identical. In simpler terms, it addresses a **difference question**: Is there a statistically significant variation in the measured outcomes across the three or more conditions experienced by the same subjects? This approach contrasts sharply with other statistical objectives, such as examining the strength and direction of association between two variables (which would require correlation analysis) or building models to forecast one variable based on others (prediction analysis, such as regression).

When the test yields a statistically significant result, it provides strong evidence against the null hypothesis, suggesting that the conditions or interventions have led to different effects on the dependent variable. However, like One-Way ANOVA, the Friedman Test is an omnibus test, meaning it only tells us that a difference exists somewhere among the groups, not precisely which pairs are significantly different from one another. Post-hoc analysis (e.g., using Wilcoxon Signed-Rank tests with Bonferroni correction) is required to pinpoint specific group differences, often involving the calculation of adjusted p-values.

Criterion 2: The Nature of Continuous Data

As previously established, the dependent measure must be either continuous (interval or ratio) or ordinal, allowing for meaningful assignment of ranks. Continuous data permits essentially infinite precision, taking on any value within a range--examples include physiological measures like heart rate, exact body measurements like height or weight, or finely scaled psychological ratings. The

ability to rank these values is fundamental to the non-parametric calculation, as the statistical power is derived from the consistent ordering within subjects.

It is equally important to recognize data types that are incompatible with the Friedman Test. These include strictly **categorical data** (nominal variables such as gender, eye color, or political affiliation), which cannot be ranked meaningfully. Similarly, **binary data** (dichotomous variables like success/failure or presence/absence of a trait) and data that are purely descriptive without inherent order are unsuitable for this rank-based comparison. Ensuring your data meets this measurement requirement prevents errors in interpretation and ensures that the ranking procedure performed by the test is statistically sound.

Criterion 3: Requirement for Three or More Conditions

The structure of the Friedman Test is specifically designed for experiments involving K geq 3 conditions or levels of measurement. It handles the simultaneous comparison of these multiple related groups on the outcome variable. This multi-group comparison efficiency is what makes it a necessary alternative to tests designed for only two groups, as it controls for the inflation of Type I error that would occur if multiple two-group comparisons were performed sequentially.

*If your research design involves only **two related groups** (e.g., pre-test vs. post-test), the appropriate non-parametric comparison test is the **Wilcoxon Signed-Rank Test**. Attempting to run a Friedman Test with only two conditions is inefficient and inappropriate, as the Wilcoxon test is optimized for that specific design and avoids unnecessary degrees of freedom in the calculation.*

Criterion 4: The Necessity of Related Samples (Repeated Measures)

The defining characteristic of the Friedman Test is its application to **related samples**, also known as repeated measures, dependent samples, or measurements within subjects. This means that each subject or block provides a measurement for every condition or group being compared. For example, if a researcher is testing the efficacy of a new diet regimen, they might measure the weight of the same cohort of participants at baseline (Group 1), after one month (Group 2), and after two months (Group 3). Since the same individual provides all three data points, the groups are inherently related.

The statistical advantage of using related samples is that it controls for individual subject variability, thereby increasing the power of the statistical test to detect genuine effects. By ranking the data within each subject, the test effectively removes the subject-specific baseline noise, focusing solely on the differences induced by the experimental conditions. This control over inter-subject variability is a key strength of the repeated measures design and the Friedman Test procedure.

Conversely, if you have three or more groups where the individuals in each group are completely

*different and unrelated (e.g., three separate treatment arms using three distinct sets of patients), then you must use the **Kruskal-Wallis One-Way ANOVA**, which is the non-parametric test for independent samples. Mixing these design types will lead to severe misinterpretation of results.*

A Practical Application of the Friedman Test

To illustrate the utility and implementation of the Friedman Test, consider a common research design in health and behavioral sciences. We will examine a scenario focused on monitoring health outcomes over time, where individual variability is high and distributional assumptions may be violated:

Scenario: A cohort of adult men is recruited via a random sample to participate in a structured, three-month comprehensive exercise program designed to improve cardiovascular health.
Repeated Measures: Data collection is scheduled at three distinct time points: Month 1 (baseline assessment), Month 2 (mid-program checkpoint), and Month 3 (final assessment).
Variable of Interest: The primary outcome measure is participants' total **Cholesterol levels** (measured in mg/dL), which is a continuous variable.

In this experimental setup, we have three related groups (the same men measured at three different times) and one dependent measure that is continuous. Prior analysis of the data reveals that the distribution of cholesterol levels is significantly **positively skewed**, failing the assumption of normality required for Repeated Measures ANOVA. Consequently, the appropriate analytical procedure is the Friedman Test. After confirming that the sample was drawn randomly and the sample size is sufficient (N > 5), we proceed with the analysis, knowing that the rank-based method will handle the non-normal distribution effectively.

Hypothesis Testing and Interpretation of Results

The process of statistical testing begins with defining the **null hypothesis** (H_0). The null hypothesis represents the default position--that the exercise program had absolutely no effect on cholesterol levels over time. Statistically, this means that the population distributions of cholesterol levels are identical across the three time points (Month 1, Month 2, and Month 3), and any observed differences are merely due to random chance or sampling error. The researchers, however, are testing the alternative hypothesis (H_A), which states that at least one of the time points differs significantly from the others, suggesting a real effect of the intervention.

When the Friedman Test is executed, the statistical software produces two key outputs: the **Chi-square statistic** (chi^2) and the corresponding **p-value**. The Chi-square statistic is a standardized measure quantifying the overall discrepancy or magnitude of difference among the average ranks of the three groups. Since the test compares rank sums, a larger Chi-square value

indicates a greater difference in cholesterol rank sums across the months, suggesting a stronger programmatic effect relative to the inherent variability within the subjects.

The p-value is the probability of observing the current data (or data more extreme) if the null hypothesis were truly correct (i.e., if the exercise program had no effect). If this p-value is calculated to be less than or equal to the predetermined significance level (typically $alpha = 0.05$), the result is considered **statistically significant**. This significance indicates that the observed difference is highly unlikely to be the result of chance alone, leading the researchers to reject the null hypothesis and conclude that the exercise program did affect cholesterol levels over the three months.

A combination of a high Chi-square statistic and a low p-value implies a significant overall effect across the time points. However, because the Friedman Test is an omnibus test, further investigation is mandatory to determine precisely which time point(s) experienced the most significant change (e.g., whether Month 3 was significantly lower than Month 1, or Month 2 was different from Month 3). This typically involves conducting follow-up post-hoc tests, such as Dunn's test, which adjust for multiple comparisons to maintain the family-wise error rate at an acceptable level.