How to Perform a Friedman Test: Definition, Formula, and Example

Authored by

stats writer

March 13, 2026

Foundations and Theoretical Framework of the Friedman Test

The **Friedman Test** represents a pivotal **non-parametric** statistical procedure used primarily to determine whether there are significant differences between three or more related groups. Originally proposed by the Nobel laureate Milton Friedman, this test functions as the non-parametric equivalent of the **repeated measures ANOVA**. It is particularly valuable when the data being analyzed does not meet the strict requirements of **normality** or homogeneity of variance, which are necessary for parametric tests to yield valid results. By focusing on the ranks of the data rather than the raw values themselves, the test provides a resilient framework for analyzing ordinal data or skewed interval data that might otherwise produce misleading outcomes in a parametric context.

In practical research environments, the **Friedman Test** is often utilized within a within-subjects design, where the same individuals or items are measured under multiple different conditions or at various points in time. This longitudinal approach allows researchers to control for individual differences between subjects, as each subject effectively serves as its own control. This increases the **statistical power** of the study by reducing the "noise" or error variance that typically arises from differences between different people in a group. Consequently, the test is a staple in fields ranging from **psychology** and **medicine** to market research and behavioral economics, where researchers frequently track the same cohort over a sequence of events.

The primary objective of this statistical method is to test the **null hypothesis**, which posits that there is no difference between the distributions of the various groups being compared. If the resulting **p-value** is lower than a predetermined threshold--usually 0.05--the null hypothesis is rejected in favor of the **alternative hypothesis**, suggesting that at least one group differs significantly from the others. This makes the **Friedman Test** an indispensable tool for identifying shifts in performance, sentiment, or physiological responses across a series of interventions or observational periods. It ensures that the conclusions drawn are statistically sound even when the data distribution is less than ideal.

Furthermore, the **Friedman Test** is highly effective when dealing with **small sample sizes** where the assumption of a normal distribution cannot be reliably tested or assumed. In such cases, **parametric** tests like the **ANOVA** may lack the robustness required to provide accurate results. By converting raw scores into ranks, the **Friedman Test** mitigates the influence of **outliers**, which can disproportionately affect the mean and variance in a parametric analysis. This makes the test a safer choice for real-world data, which is frequently messy and contains extreme values that do not reflect the general trend of the population.

Key Scenarios and Use Cases in Research

The **Friedman Test** is commonly applied in two distinct research scenarios that involve repeated measures. The first scenario involves measuring the mean scores of subjects across three or more specific time points. For instance, a sports scientist might want to track the resting heart rate of athletes at various intervals: one month before starting a high-intensity **training program**, one month into the program, and two months after the program has concluded. By using the **Friedman Test**, the scientist can determine if the training intervention led to a **statistically significant** change in cardiovascular health across these three distinct periods.

The second major application is measuring the scores of subjects under three or more different conditions or treatments. Imagine a **user experience** study where participants are asked to watch three different types of cinematic content and rate their level of engagement or enjoyment for each. Because each individual participant provides a rating for every movie, the samples are "related" or "matched." The **Friedman Test** allows the researcher to see if there is a genuine preference for one movie over the others, or if the differences in ratings are simply due to **random chance**. This ability to compare multiple conditions within the same subject pool is what makes the test so versatile across different scientific disciplines.

Another example can be found in **clinical trials** where the same group of patients is exposed to different medications or dosages to observe the varying effects on a particular symptom. For example, a clinician might test three different concentrations of a drug to find the optimal balance between efficacy and side effects. Since the same patients are used for each dosage level to maintain consistency, the **Friedman Test** is the appropriate tool to analyze the resulting data. It helps in identifying whether the dosage level has a significant impact on the patient's recovery rate, ensuring that the clinical findings are backed by rigorous statistical evidence.

Finally, the test is frequently used in **educational research** to evaluate the effectiveness of different teaching methodologies. A teacher might implement three different instructional strategies over a semester and measure student performance after each module. By applying the **Friedman Test** to the students' grades, the educator can objectively determine which method yielded the best learning outcomes. This data-driven approach to education allows for the continuous refinement of pedagogical techniques, ultimately benefiting the students' academic growth and the institution's overall effectiveness.

Core Assumptions and Data Requirements

Before performing a **Friedman Test**, it is essential to ensure that the data meets certain underlying assumptions. While it is a **non-parametric** test and therefore more flexible than its parametric counterparts, it still requires specific conditions to be met for the results to be valid. The first assumption is that the **dependent variable** should be measured at the **ordinal** or **continuous** level. This means the data must be capable of being ranked. Examples include **Likert scales**, time

in seconds, or scores on a standardized test. If the data is purely categorical without any inherent order, this test cannot be applied.

The second critical assumption is that the **independent variable** must consist of three or more related groups or "blocks." These groups are typically the same subjects being measured multiple times or subjects that have been matched based on specific characteristics. This relationship between the groups is the defining feature of the **Friedman Test**. If the groups were independent--meaning different people were in each group--a different test, such as the Kruskal-Wallis test, would be required instead. Understanding the relationship between your samples is vital for choosing the correct statistical path.

The third assumption involves the **random sampling** of subjects from the population. To generalize the findings of the **Friedman Test** to a larger group, the participants in the study should be selected randomly. Furthermore, while the test does not require a normal distribution, it does assume that the distributions of the various groups have a similar shape. If one group is heavily skewed to the left while another is skewed to the right, the test's ability to accurately compare the medians may be compromised. Ensuring these basic criteria are met allows the researcher to proceed with confidence in their statistical analysis.

Lastly, it is important to note that the **Friedman Test** assumes that the observations within each block are independent of each other, except for the relationship defined by the groups. For example, in a study involving ten patients, the reaction time of Patient A should not influence the reaction time of Patient B. Violation of this **independence of observations** can lead to an increased risk of **Type I errors**, where a researcher might incorrectly conclude that a significant difference exists when it does not. Maintaining a clean experimental design is therefore just as important as the mathematical calculation itself.

The Mathematical Logic and Formula

The mathematical foundation of the **Friedman Test** relies on the ranking of data across the different conditions for each individual subject. Instead of looking at the raw values, the test looks at how those values compare to one another within each row. For each subject, the scores are assigned a rank of 1, 2, 3, and so on, based on their magnitude. If there are three groups, the lowest score for a subject gets a rank of 1, the middle score a rank of 2, and the highest a rank of 3. This process is repeated for every subject in the dataset, effectively standardizing the data across the blocks.

The test statistic, often denoted as **Q** or **Friedman's chi-square**, is calculated using the sum of these ranks. The formula incorporates the number of subjects (n), the number of groups or conditions (k), and the sum of the ranks for each group. By comparing the actual distribution of these rank sums to the distribution expected under the **null hypothesis**, the test determines if the

observed differences are statistically significant. The formula is designed to produce a value that follows a chi-square distribution with k-1 **degrees of freedom**.

One of the unique aspects of this formula is its ability to handle "ties" in the data. If a subject has the same score in two different conditions, they are assigned an average rank. For example, if two scores are tied for the 1st and 2nd positions, both are given a rank of 1.5. While heavy ties can slightly reduce the **statistical power** of the test, modern software packages include adjustments in the formula to account for this, ensuring that the resulting **p-value** remains as accurate as possible. This robustness is a key reason why the test is preferred for **ordinal data** where ties are common.

Understanding the formula also highlights why the **Friedman Test** is less sensitive to extreme **outliers** than the **repeated measures ANOVA**. Because a massive outlier only receives the highest rank (e.g., a rank of 3 in a 3-group study), its specific numerical value does not disproportionately pull the group mean. This "ranking transformation" effectively dampens the noise caused by anomalous data points, focusing instead on the consistent patterns of "better" or "worse" across the conditions. It is this focus on relative performance that makes the test a powerful tool for ranking-based analysis.

Illustrative Example: Drug Reaction Times

To better understand the application of this test, let us consider a practical example involving pharmacology. Suppose a group of researchers wants to determine if the mean reaction time of subjects varies significantly when they are administered three different drugs. To conduct this experiment, they recruit 10 patients and measure each individual's reaction time, recorded in seconds, after taking Drug A, Drug B, and Drug C. This design ensures that the samples are related, as each patient contributes a data point to every drug category. The raw data collected from this study is organized into a table for analysis.

Patient	Drug 1	Drug 2	Drug 3
Patient 1	4	5	2
Patient 2	6	6	4
Patient 3	3	8	4
Patient 4	4	7	3
Patient 5	3	7	2
Patient 6	2	8	2
Patient 7	2	4	1
Patient 8	7	6	4
Patient 9	6	4	3
Patient 10	5	5	2

In this specific dataset, we can observe the varying reaction times for each of the ten patients. For instance, Patient 1 might have a reaction time of 0.5 seconds on Drug A, 0.7 seconds on Drug B, and 0.6 seconds on Drug C. Before the **Friedman Test** is performed, these values must be ranked within the row for Patient 1. In this case, Drug A would receive a rank of 1, Drug C a rank of 2, and Drug B a rank of 3. This ranking process is meticulously applied to all ten patients, creating a new matrix of ranks that will serve as the basis for the **statistical significance** calculation.

The goal of this study is to move beyond mere observation and determine if the differences seen in the table are consistent enough across the 10 patients to be considered a general effect of the drugs. Without a formal test like the **Friedman Test**, it would be difficult to say whether Drug B is truly "slower" than Drug A, or if the variation we see is just **random noise**. By utilizing this **non-parametric** approach, the researchers can account for the fact that some patients naturally have faster reaction times than others, focusing purely on how the drugs shift those baseline speeds.

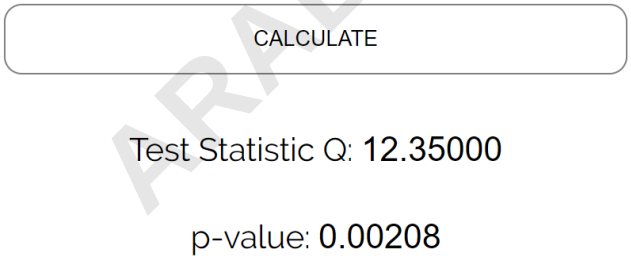
Executing the Friedman Test: Step-by-Step

The first step in any statistical analysis is to clearly state the **hypotheses**. For our drug reaction time example, the **null hypothesis (H0)** is that the mean reaction times across the populations are all equal ($\mu1 = \mu2 = \mu3$). This suggests that the type of drug has no impact on reaction time. Conversely, the **alternative hypothesis (Ha)** states that at least one population mean is different from the rest. This sets the stage for the calculation, providing a clear question that the **Friedman Test** will answer through mathematical evidence.

In the second step, the actual **Friedman Test** calculation is performed. In modern research, this is typically done using statistical software or specialized calculators. The ranked data is fed into the algorithm, which computes the sum of ranks for each drug and applies the **Friedman formula** to generate the **test statistic (Q)**. This value represents the magnitude of the difference between the groups. Below is an example of the input interface used to process the reaction time data for our ten patients.

Group 1	Group 2	Group 3	Group 4	Group 5
4	5	2		
6	6	4		
3	8	4		
4	7	3		
3	7	2		
2	8	2		
2	4	1		
7	6	4		
6	4	3		
5	5	2		

Once the calculation is triggered, the software compares the calculated **Q** value against the critical value from a chi-square distribution. The output will typically display the test statistic, the **degrees of freedom**, and the **p-value**. The **p-value** is the most critical component for interpretation, as it tells us the probability of obtaining these results if the **null hypothesis** were actually true. An example of the final output for our reaction time study is shown in the image below.



Interpreting the Statistical Results

Interpreting the results of the **Friedman Test** requires a look at the specific values generated in the output. In our example, the test statistic is **Q = 12.35**, and the corresponding **p-value** is **0.00208**. In the world of statistics, a common threshold for significance is **alpha = 0.05**. Since our **p-value** of 0.00208 is significantly lower than 0.05, we have strong evidence to reject the **null hypothesis**.

This indicates that the differences in reaction times between the three drugs are not due to chance, but are **statistically significant**.

Rejecting the null hypothesis allows us to conclude that the type of drug administered has a definitive effect on the reaction times of the patients. However, it is important to note that the **Friedman Test** is an "omnibus" test. This means it tells us that a difference exists somewhere among the groups, but it does not specify exactly which groups are different from each other. For example, it doesn't tell us if Drug A is different from Drug B, or if Drug B is different from Drug C. To find those specific differences, researchers must perform **post-hoc tests**.

Common **post-hoc tests** used following a significant **Friedman Test** include the **Wilcoxon signed-rank test** with a **Bonferroni correction**. This follow-up analysis allows for pairwise comparisons between the drugs while controlling for the increased risk of errors that comes with performing multiple tests. By conducting these additional steps, researchers can gain a more nuanced understanding of their data, identifying exactly which intervention is the most effective or which time point showed the most significant change.

Formal Reporting and Documentation

The final stage of the process is reporting the results in a clear and professional manner, suitable for an academic journal or a business report. A good report should include the purpose of the test, the sample size, the conditions being compared, the test statistic, and the **p-value**. It should also state whether the results were **statistically significant** and what that means in the context of the research question. Transparency in reporting ensures that other researchers can evaluate the validity of the study and replicate the findings if necessary.

Following our drug reaction time example, a formal report might look like this: "A **Friedman Test** was conducted on 10 patients to examine the effect that three different drugs had on response time. Each patient was measured once for each drug to ensure a **repeated measures design**. The results of the analysis showed that the type of drug used led to **statistically significant** differences in response time ($Q = 12.35$, $p = 0.00208$). Consequently, we reject the null hypothesis and conclude that the medication significantly impacts patient reaction speed."

This style of reporting provides all the necessary technical details while remaining accessible to the reader. It avoids unnecessary jargon and focuses on the core findings of the statistical analysis. By adhering to these reporting standards, researchers contribute to a body of knowledge that is rigorous, verifiable, and meaningful. Whether you are writing for a medical board or a university thesis, the clarity of your statistical reporting is just as important as the data itself.

Software Implementation and Further Learning

In the modern era, researchers rarely perform the **Friedman Test** by hand. Instead, they rely on powerful programming languages and statistical software to handle the data. Python is one of the most popular choices for this task, thanks to libraries like **SciPy** and **Statsmodels**, which contain built-in functions for non-parametric testing. Using code ensures that the analysis is reproducible and can be easily updated if more data is collected in the future.

For those interested in learning the technical implementation of this test, there are many high-quality resources available online. These tutorials often provide step-by-step guides on how to structure your data in a **Pandas DataFrame**, how to call the **friedmanchisquare** function, and how to interpret the resulting objects. Mastering these tools is essential for any modern data scientist or researcher looking to perform high-level statistical analysis. You can find a detailed guide on this specific implementation here:

How to Perform the Friedman Test in Python

Beyond **Python**, other software packages like **R**, **SPSS**, and **SAS** also offer robust support for the **Friedman Test**. Each of these tools has its own strengths, but they all follow the same underlying logic of ranking and chi-square approximation. By expanding your toolkit to include these various platforms, you become more versatile as a researcher, capable of handling diverse datasets and meeting the specific requirements of different scientific disciplines. Continuous learning in the field of **statistics** is the key to unlocking deeper insights from your data.