Learn to Perform the Nemenyi Post-Hoc Test with Python

Authored by

Mohammed looti

November 5, 2025

The Necessity of Non-Parametric Post-Hoc Analysis

The **Nemenyi test** is an indispensable tool in statistical inference, serving as a robust non-parametric equivalent to procedures like the Repeated Measures ANOVA. This test is specifically designed for situations where researchers have measured the same subjects under three or more distinct conditions (a classic repeated measures design) but where the underlying data distributions violate the stringent assumptions required by parametric tests, such as normality or sphericity. The Nemenyi procedure allows for detailed examination of whether statistically significant differences exist among the central tendencies of these related groups.

It is crucial to understand that the Nemenyi procedure is a **post-hoc test**, meaning it cannot be applied in isolation. It must follow a significant result from an omnibus test that first confirms an overall effect. For non-parametric repeated measures data, this prerequisite test is the Friedman Test. If the resulting p-value from the Friedman Test falls below the predetermined significance threshold (commonly $p < 0.05$), we reject the null hypothesis, confirming that differences exist somewhere among the group distributions.

Only after the overall significance has been established can we proceed to the Nemenyi post-hoc test. This powerful follow-up step conducts all possible pairwise comparisons between the groups. Importantly, the Nemenyi method incorporates adjustments to control the Family-Wise Error Rate (FWER). This control mechanism is vital because it minimizes the risk of incorrectly declaring a difference significant (Type I error) when running multiple simultaneous comparisons, thereby ensuring the reliability and high confidence of the final conclusions.

The Foundation: Why the Friedman Test is Necessary

The selection of the Friedman Test is appropriate precisely when the stringent assumptions of parametric methods are violated, or when the data collected are inherently ordinal. As a rank-based test, the Friedman procedure is less sensitive to outliers and skewness than its parametric counterparts, making it highly valuable for experimental designs involving multiple, linked measurements on the same collection of subjects or blocks. It transforms the raw data into ranks within each subject, thus evaluating the consistency of these ranks across the different treatment conditions.

In rigorous statistical language, the Friedman Test assesses two competing hypotheses regarding the underlying population distributions:

The Null Hypothesis (H0): This hypothesis posits that the distributions of the measurements across all groups or conditions are statistically identical. In practical terms, it suggests that the treatment or experimental condition has produced no discernible effect on the outcome measure.

The Alternative Hypothesis (Ha): This hypothesis states that a difference exists, specifically that at least one population distribution is stochastically different from the others. This implies that the treatment or condition has had a measurable impact.

A significant outcome from the omnibus test, however, offers limited insight; it merely confirms that **at least one pair** of groups contributes to the overall effect. It fails to identify which specific conditions are responsible for this variation. This critical ambiguity mandates the subsequent use of a specialized post-hoc test, such as the Nemenyi procedure, to systematically isolate and identify the specific sources of statistical difference.

Practical Example: Setting Up the Data in Python

To clearly demonstrate this analytical approach, let us consider a hypothetical research scenario focused on evaluating the efficacy of three distinct pharmaceutical drugs on patient reaction times. In this study, a cohort of 10 patients participates, and each patient is tested under all three drug conditions sequentially. The dependent variable is the reaction time, measured in seconds. This structure perfectly aligns with the requirements of a classic repeated measures design, where the data points within a condition are related through the subjects.

The primary research objective is to determine if the average reaction times differ significantly across the three drug conditions. The data must be prepared such that the scores for a specific drug are contained within a single list or array, and the positional index within these arrays must correspond consistently to the same subject across all conditions.

We translate the measured response times for our 10 patients across the three drug trials into three separate Python arrays. It is mandatory that these arrays possess identical lengths, as each index pair represents observations from the same unique subject:

group1 =
group2 =
group3 =

Initial Analysis: Executing and Interpreting the Friedman Test

The essential first step in the analysis sequence is the execution of the overall Friedman Test. In the Python environment, this is efficiently achieved using the `friedmanchisquare` function, which is readily available within the scipy.stats library. SciPy is a core foundation for scientific computing and provides reliable, validated statistical functions.

We begin by importing the required module and executing the test, passing all three arrays containing our drug trial data as positional arguments to the function:

from scipy import stats

#perform Friedman Test
stats.friedmanchisquare(group1, group2, group3)

FriedmanchisquareResult(statistic=13.3513513, pvalue=0.00126122012)

The resulting output furnishes two critical pieces of information: the computed test statistic (chi^2) and its corresponding p-value. In this particular analysis, the calculated test statistic is approximately **13.35**, and the associated p-value is extremely low, at **0.00126**. Since the observed p-value (0.00126) is substantially smaller than the conventional significance level of $alpha = 0.05$, we have compelling statistical evidence to **reject the null hypothesis (H0)**. This robust finding confirms that the administration of different drugs results in statistically significant differences in patient reaction times. However, the analysis remains incomplete until we identify which specific drug comparisons drive this overall effect.

Transitioning to the Nemenyi Procedure

Given the significant result from the Friedman Test, the subsequent mandatory step is to employ a suitable post-hoc test to precisely locate the pair-wise differences. The Nemenyi procedure is the statistically recommended method following a significant Friedman result. Its mechanism involves calculating the absolute difference between the mean rank sums for every possible pair of groups. This difference is then compared against a calculated critical range, which is determined by factoring in the number of groups and the number of subjects. This methodology ensures that the Family-Wise Error Rate remains controlled throughout the multiple comparisons.

To seamlessly execute the Nemenyi procedure within Python, we must utilize the `scikit-posthocs` library. This external library offers comprehensive implementations of various post-hoc tests suitable for both parametric and non-parametric data structures. If this library is not currently installed in your environment, you must first install it using the standard package installer, `pip`:

pip install scikit-posthocs

Once the necessary library has been successfully installed and imported, the data must be rigorously structured to meet the specific input requirements of the pairwise comparison function, ensuring accurate analysis.

Implementing the Nemenyi Test with scikit-posthocs

The Nemenyi test is performed by invoking the `posthoc_nemenyi_friedman()` function from the

`scikit-posthocs` library. A critical prerequisite for this function is the specific organization of the input data: the rows of the input matrix must represent the subjects (or blocks), and the columns must represent the distinct treatments (groups).

Since our initial data structure consisted of three separate Python arrays (`group1`, `group2`, `group3`) defined as column vectors, we must first consolidate them into a single structure, typically a NumPy array. Crucially, we must then transpose this resulting matrix using the `data.T` operation. This transposition ensures that the 10 subjects are correctly aligned vertically across the rows, fulfilling the function's structural requirement.

import scikit_posthocs as sp
import numpy as np

#combine three groups into one array
data = np.array()

#perform Nemenyi post-hoc test (Note the transposition: data.T)
sp.posthoc_nemenyi_friedman(data.T)

0 1 2
0 1.000000 0.437407 0.065303
1 0.437407 1.000000 0.001533
2 0.065303 0.001533 1.000000

This step, particularly the transposition of the NumPy array, is non-negotiable for correct execution. The resulting matrix structure now aligns the 10 patients as rows and the 3 drug conditions as columns (labeled 0, 1, and 2).

Interpreting Pairwise Comparisons

The output generated by the Nemenyi post-hoc test is presented as a symmetrical matrix. This matrix concisely displays the adjusted p-values for every possible pairwise comparison among the three drug groups. The columns and rows are numerically labeled 0, 1, and 2, which correspond directly to our input variables `group1`, `group2`, and `group3`, representing Drug 1, Drug 2, and Drug 3, respectively.

To interpret these results, we compare each adjusted p-value against our predefined significance level, $alpha = 0.05$. If an adjusted p-value is less than 0.05, we confidently conclude that the difference between that specific pair of groups is statistically significant, after controlling for multiple testing errors.

We extract the critical p-values from the matrix:

P-value for Drug 1 (Group 0) vs. Drug 2 (Group 1): **0.4374**

P-value for Drug 1 (Group 0) vs. Drug 3 (Group 2): **0.0653**

P-value for Drug 2 (Group 1) vs. Drug 3 (Group 2): **0.0015**

Applying the $alpha = 0.05$ threshold allows us to draw precise conclusions regarding the efficacy of the drugs:

The comparison between Drug 1 and Drug 2 (p = 0.4374) is definitively **not statistically significant**.

The comparison between Drug 1 and Drug 3 (p = 0.0653) is **not statistically significant** at the standard 0.05 level, although it approaches the threshold closely.

The comparison between Drug 2 and Drug 3 (p = 0.0015) is **highly statistically significant** (p < 0.05).

Based on the comprehensive results of the Nemenyi test, we can confidently assert that the only two drugs exhibiting a statistically significant difference in patient reaction times are Drug 2 (Group 1) and Drug 3 (Group 2). The research conclusion should emphasize that switching between Drug 2 and Drug 3 significantly alters patient reaction speed, while differences involving Drug 1 cannot be distinguished reliably from the others at the chosen level of significance.