Which test should you use: Tukey, Bonferroni or Scheffe?

Authored by

stats writer

December 14, 2025

When conducting advanced statistical analysis involving comparisons across multiple groups, researchers often face the critical decision of selecting the appropriate post-hoc test. This choice is vital for accurately interpreting results while maintaining control over potential analytical errors. This guide focuses on three of the most widely utilized multiple comparison procedures--the **Tukey HSD test**, the **Bonferroni correction**, and **Scheffe's method**--detailing their specific use cases and underlying assumptions. Choosing the wrong method can significantly inflate the likelihood of false positives, making an understanding of these differences paramount for robust scientific reporting.

The Necessity of Post-Hoc Analysis

The journey into multiple group comparisons typically begins with a global test, such as a one-way ANOVA, designed to detect general differences across treatment conditions. A one-way ANOVA is fundamentally used to determine whether or not there is a statistically significant difference among the means of three or more independent groups. It serves as an initial gatekeeper for the data analysis process, indicating the presence of variation across the groups being studied.

Once the ANOVA test is performed, the core output is the overall F-statistic and its associated p-value. If this overall p-value falls below a predetermined significance level (e.g., 0.05), we can confidently state that sufficient evidence exists to reject the null hypothesis--meaning that at least one group mean differs from the others. However, this finding is non-specific; it is an omnibus result that does not identify **where** the differences lie. It only confirms that the collection of group means are not all equal.

To pinpoint the exact groups that differ statistically from one another, we must perform a secondary analysis. This is where post-hoc tests become indispensable. These procedures are specifically designed to conduct multiple comparisons while rigorously controlling the family-wise error rate (FWER)--the probability of making at least one Type I error across the entire set of comparisons. Without this control, the chance of falsely declaring a difference increases dramatically as the number of comparisons grows.

Three of the most reliable and commonly employed post-hoc procedures are:

The **Tukey HSD Method** (Honestly Significant Difference)
The **Scheffe Method**
The **Bonferroni Method**

This tutorial provides a comprehensive overview of each method along with instructions on which post-hoc test is appropriate depending on the specific research question and study design constraints.

The Tukey HSD Method

The Tukey post-hoc test, often referred to as Tukey's Honestly Significant Difference (HSD) test, is the preferred choice for general comparison among multiple groups when the research goal is to examine all possible pairwise comparisons. The term "pairwise" explicitly means that we are only comparing two group means at a time. This method is specifically designed to hold the family-wise error rate at the specified alpha level across all these comparisons simultaneously, making it robust for exploratory analyses where the researcher is interested in every potential mean difference.

The Tukey HSD test operates under the assumption that the sample sizes for each group are equal, characterizing a balanced design. If the group sample sizes are unequal (an unbalanced design), researchers should utilize a slight modification of the procedure known as the **Tukey-Kramer test**. Both versions effectively control the overall error rate, offering a strong balance between controlling Type I errors and maintaining adequate statistical power compared to more conservative methods.

For example, suppose a study investigates three distinct treatment groups--A, B, and C. The Tukey post-hoc test systematically facilitates all possible two-group comparisons that must be evaluated:

Comparison 1: Testing if μA is equal to μB
Comparison 2: Testing if μA is equal to μC
Comparison 3: Testing if μB is equal to μC

It is important to note that for a study involving *k* number of groups, the total count of possible pairwise comparisons is calculated using the formula $k(k-1)/2$. Because the Tukey method is specifically engineered to handle this high volume of simultaneous comparisons efficiently, it is the standard recommendation when the research objective is exploratory and includes no a priori hypotheses about specific pairs.

Understanding Scheffe's Method

The Scheffe post-hoc test provides the greatest flexibility among these three methods because it is capable of testing *all* possible contrasts between group means, not just simple pairwise comparisons. A contrast is a linear combination of means where the coefficients sum to zero, allowing researchers to compare more than two means simultaneously or evaluate complex hypotheses involving weighted averages of the groups. This extensive capability makes it suitable for complex experimental designs or when unexpected differences are observed post-experimentally.

For instance, in a four-group study (A, B, C, D), Scheffe's method allows us to construct and test

complex comparisons such as:

A complex comparison of differences: μA - μB = μC - μD
A complex comparison of averages: μA + μD = μB + μC

While the flexibility of Scheffe's method is highly advantageous for complex or unanticipated findings, it is important to understand its trade-off. Scheffe's method is recognized as the most statistically conservative of these three tests. This conservatism results in the widest confidence intervals for the observed differences between means, especially in simple pairwise testing scenarios. Consequently, it possesses the lowest statistical power, meaning it has the lowest ability to detect true differences between the groups compared to Tukey or Bonferroni, making it harder to achieve statistical significance. However, a major practical advantage is that Scheffe's procedure is robust and can be reliably used whether or not the group sample sizes are equal.

Implementing the Bonferroni Correction

The Bonferroni post-hoc test employs a different approach and is ideally suited when the researcher has a specific, limited set of **planned comparisons** that were formulated before any data analysis took place. This is a confirmatory approach, contrasting sharply with the exploratory nature of the Tukey test. Bonferroni only adjusts for the number of comparisons the researcher explicitly intends to make, making it highly efficient when those comparisons are few.

The core mechanism of the Bonferroni correction involves controlling the FWER by adjusting the significance level (alpha) for each individual test. If *c* represents the total number of planned comparisons, the critical alpha level for each test is recalculated as α/c. This stringent adjustment ensures that the overall probability of committing a Type I error across the entire set of tests remains at or below the original alpha level.

For example, suppose we have three groups--A, B, C--and based on existing literature, we only hypothesize differences in the following two pairings:

Planned Comparison 1: Testing if μA is equal to μB
Planned Comparison 2: Testing if μB is equal to μC

When the analysis is restricted to a small, pre-defined set of planned comparisons, the Bonferroni correction generally produces the most narrow confidence intervals compared to both Tukey and Scheffe. These narrower intervals translate directly into higher statistical power, granting it the greatest ability to detect true differences among the specific groups of interest. Furthermore, like Scheffe's method, the Bonferroni correction is versatile and can be used effectively whether or not the group sample sizes are equal.

Decision Framework: Which Test to Use?

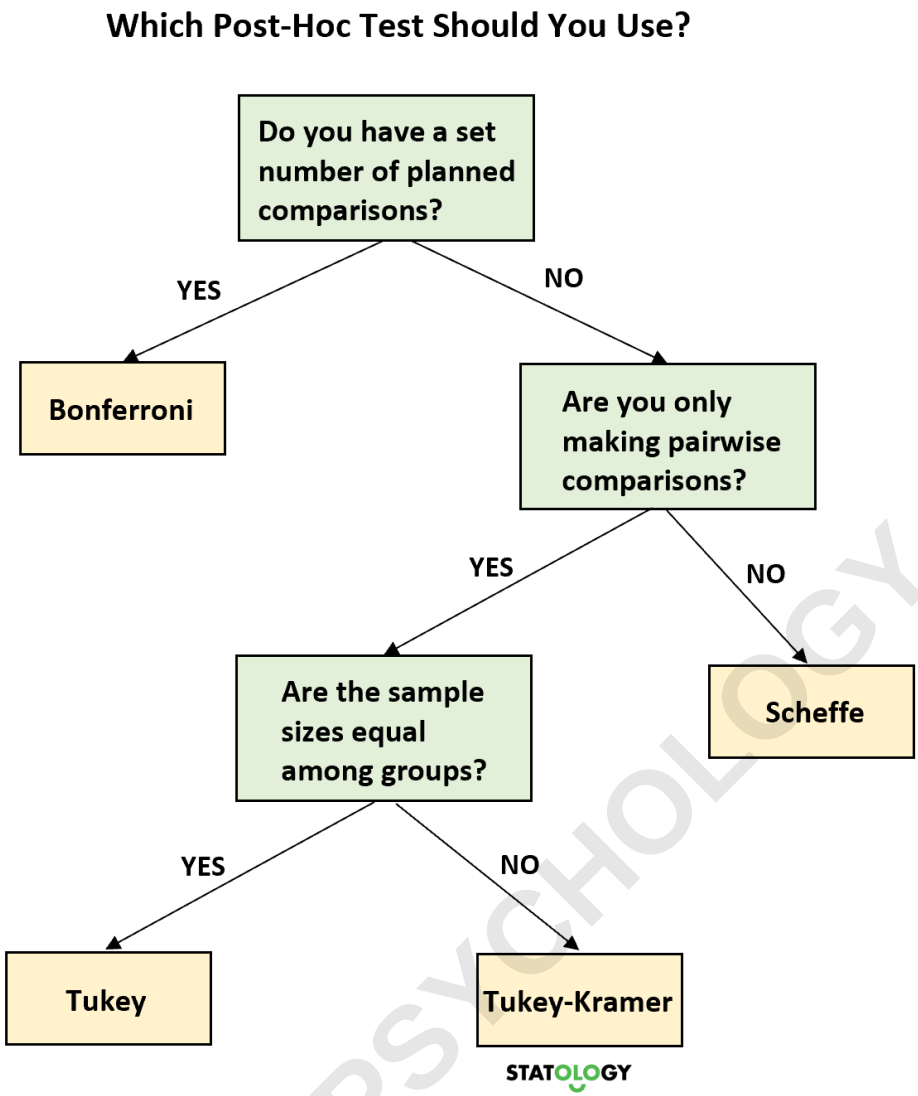
The selection of the appropriate post-hoc test should be determined by the researcher's intent (exploratory vs. confirmatory) and the complexity of the contrasts required. The key distinction lies in whether the comparisons are predetermined and limited, or if all possible comparisons must be evaluated.

The following summarized points guide the selection:

Tukey's HSD: Recommended for **exploratory research** that mandates testing all possible simple pairwise comparisons. It is the preferred general-use method when sample sizes are equal, offering good power and strong FWER control.
Bonferroni: The most powerful choice for **confirmatory research** when only a small number of specific comparisons are **planned** ahead of time. It becomes overly conservative and suffers a loss of power if the number of comparisons is large.
Scheffe's: Utilized primarily for testing **complex contrasts** or linear combinations of means. It is the most conservative option, resulting in lower statistical power, but guarantees FWER control for the largest possible set of hypotheses.

Visualizing the Decision Process

To further assist researchers in making an informed choice, the following decision tree provides a visual representation based on common scenarios encountered in ANOVA follow-up testing. The framework highlights the optimal statistical procedure based on whether the comparisons are planned or data-driven.



Closing Thoughts on Research Integrity

A crucial ethical requirement in statistical reporting is that the choice of the specific post-hoc test must be firmly established and documented **before** conducting the experiment and initiating data analysis. This commitment to a predetermined analytical plan is essential for maintaining research integrity.

Failing to choose the method beforehand and instead selecting a test after observing the initial results increases the risk of "p-hacking"--the dishonest practice of choosing the analytical approach that is most likely to produce spurious significant results. Such practices artificially inflate findings and undermine the validity of scientific conclusions. Therefore, the decision regarding Bonferroni (planned, focused), Tukey (exploratory, pairwise), or Scheffe (complex, conservative) must align with the foundational research hypotheses.

In practice, researchers benefit greatly from technological advances, as most modern statistical software packages are fully capable of performing these post-hoc tests automatically. This convenience allows the statistical community to focus less on manual computation and more on the appropriate application and interpretation of the results based on their predetermined methodology.