Learning Post-Hoc Pairwise Comparisons After ANOVA in R

Authored by

Mohammed looti

November 1, 2025

The **Analysis of Variance (ANOVA)** is a foundational statistical procedure employed to ascertain whether meaningful variation exists among the means of three or more independent populations. In the context of experimental research, the ANOVA serves as the essential omnibus test, providing an initial determination of whether the treatment effects are collectively consequential.

When conducting a one-way ANOVA, the data is evaluated against a set of competing hypotheses:

H0: The **Null Hypothesis** asserts that all population group means are equivalent, implying that the independent variable has no effect.
HA: The **Alternative Hypothesis** posits that at least one group mean is different from the others, suggesting a genuine effect.

A positive result--where the overall **p-value** is below the predefined significance threshold (commonly α **= .05**)--compels us to reject the null hypothesis. While this confirms that a **statistically significant difference** exists among the means, the ANOVA itself fails to pinpoint **which** specific pairs are driving this observed effect.

To obtain this necessary granularity, researchers must proceed with **post-hoc pairwise comparisons**. These subsequent tests are critical because they implement crucial adjustments to the p-values, thereby controlling for the inflated risk of committing a Type I error (false positive) that inherently arises from performing multiple comparisons simultaneously.

This guide details the practical execution of several powerful and widely used post-hoc methods within the **R statistical environment**:

The **Tukey Method** (Tukey's Honestly Significant Difference, HSD)
The **Scheffe Method**
The **Bonferroni Method**
The **Holm Method** (Holm-Bonferroni Procedure)

Setting Up and Executing the Initial ANOVA in R

To demonstrate the workflow for conducting post-hoc analyses, we utilize a common experimental scenario. Imagine a researcher investigating the effectiveness of three different study methods on final exam scores. Thirty students are equally divided and randomly assigned to one of three groups (n=10 per group). The resulting exam scores form a balanced dataset suitable for a **one-way ANOVA**.

The initial procedural step requires performing the omnibus ANOVA test to ascertain if the independent variable (studying technique) significantly impacts the dependent variable (exam score). In **R**, we first structure the data frame and then apply the standard **aov()** function, which is

designed for fitting analysis of variance models:

```
#create data frame
df <- data.frame(technique = rep(c("tech1", "tech2", "tech3"), each=10),
score = c(76, 77, 77, 81, 82, 82, 83, 84, 85, 89,
81, 82, 83, 83, 83, 84, 87, 90, 92, 93,
77, 78, 79, 88, 89, 90, 91, 95, 95, 98))

#perform one-way ANOVA
model <- aov(score ~ technique, data = df)

#view output of ANOVA
summary(model)
```

```
Df Sum Sq Mean Sq F value Pr(>F)
technique 2 211.5 105.73 3.415 0.0476 *
Residuals 27 836.0 30.96
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting ANOVA table is inspected for the significance of the `technique` factor. We observe that the overall **p-value** is 0.0476. Because this value falls below the conventional alpha level of 0.05, we confidently reject the null hypothesis, confirming the existence of at least one **statistically significant difference** among the mean exam scores.

The rejection of the null hypothesis in the omnibus test necessitates the transition to **post-hoc comparisons**. This next critical phase is dedicated to dissecting the overall effect by systematically comparing every possible pair of group means while rigorously controlling the family-wise error rate (FWER) across all comparisons.

Implementing the Tukey HSD Method

The **Tukey Honestly Significant Difference (HSD) method** is widely regarded as the most robust and powerful post-hoc test, particularly when dealing with **ANOVA** designs where sample sizes are equal across all groups (a balanced design), as is the configuration in our current example. The HSD procedure compares every possible pair of means using the studentized range distribution, which offers optimal power while strictly controlling the family-wise error rate (FWER).

Controlling the FWER--the probability of incorrectly rejecting at least one true null hypothesis in the family of comparisons--is the primary objective of any post-hoc test. In **R**, the implementation is streamlined using the base function **TukeyHSD()**, which requires the fitted ANOVA model object

as its primary argument:

#perform the Tukey post-hoc method
TukeyHSD(model, conf.level=.95)

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = score ~ technique, data = df)

$technique
diff lwr upr p adj
tech2-tech1 4.2 -1.9700112 10.370011 0.2281369
tech3-tech1 6.4 0.2299888 12.570011 0.0409017
tech3-tech2 2.2 -3.9700112 8.370011 0.6547756

The output matrix provides the difference in means (`diff`), the confidence interval bounds (`lwr` and `upr`), and the adjusted p-value (`p adj`) for each comparison. Upon review, only the contrast between `tech3` and `tech1` yields an adjusted **p-value** of 0.0409, which successfully falls below the 0.05 significance threshold. Further confirmation is derived from the confidence interval (0.2299 to 12.5700), which emphatically does not contain zero.

Consequently, based on the **Tukey Method**, we conclude that students utilizing Technique 3 achieved statistically significantly higher exam scores compared to those utilizing Technique 1. The differences involving Technique 2 were not significant.

Evaluating Results Using the Conservative Scheffe Method

The **Scheffe Method** is inherently the most conservative of the common post-hoc tests. While it offers the strongest assurance regarding the control of the family-wise error rate for all possible contrasts, including complex ones, its primary drawback is its reduced **statistical power** for simple pairwise comparisons. This conservatism increases the likelihood of committing a Type II error--failing to detect a true difference (false negative).

Researchers typically reserve the Scheffe method for situations involving unbalanced designs (unequal sample sizes) or when their primary interest lies in exploring non-standard, complex contrasts (e.g., comparing the average of two groups against a third group). For simple pairwise comparisons in a balanced design, it is often overly strict.

In **R**, the Scheffe test is not part of the base installation and requires the **DescTools** package, which provides the dedicated **ScheffeTest()** function. We load the library and then apply the test to

our existing ANOVA model:

library(DescTools)

#perform the Scheffe post-hoc method
ScheffeTest(model)

Posthoc multiple comparisons of means: Scheffe Test
95% family-wise confidence level

$technique
diff lwr.ci upr.ci pval
tech2-tech1 4.2 -2.24527202 10.645272 0.2582
tech3-tech1 6.4 -0.04527202 12.845272 0.0519 .
tech3-tech2 2.2 -4.24527202 8.645272 0.6803

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 156

As anticipated given its conservatism, the adjusted **p-value** for the `tech3-tech1` comparison is 0.0519. This result marginally exceeds the 0.05 threshold. Consequently, under the strict criteria of the **Scheffe Method**, we fail to reject the null hypothesis for all pairs, leading to the conclusion that no pair exhibits a **statistically significant difference** in mean performance. This illustrates the trade-off between strict FWER control and statistical power.

Analyzing Comparisons Using the Bonferroni Correction

The **Bonferroni Method** represents a straightforward and classical technique for managing the family-wise error rate. Its mechanism is inherently simple: it divides the original alpha level by the total number of comparisons (m) to create a new, stricter threshold for significance. This method is generally recommended when the researcher has a small, predefined set of comparisons, rather than exhaustively testing all possible pairs.

Although highly transparent, the Bonferroni correction is known for its tendency toward excessive conservatism, especially when many comparisons are involved. This severity often results in a significant loss of **statistical power**, mirroring the drawback observed in the Scheffe test, thereby increasing the risk of Type II errors.

In **R**, the Bonferroni adjustment is easily applied using the versatile **pairwise.t.test()** function. We specify the `p.adj` parameter as `'bonferroni'` to implement the correction on the t-tests performed between all pairs:

```
#perform the Bonferroni post-hoc method
pairwise.t.test(df$score, df$technique, p.adj='bonferroni')
```

Pairwise comparisons using t tests with pooled SD

data: df$score and df$technique

tech1 tech2
tech2 0.309 -
tech3 0.048 1.000

P value adjustment method: bonferroni

The resulting matrix confirms that the adjusted p-value for the difference between Technique 1 and Technique 3 is 0.048. Since this value is just below the 0.05 threshold, it warrants rejection of the null hypothesis for this specific pair.

Following the criteria established by the **Bonferroni Method**, we affirm the conclusion that the only statistically significant difference in mean exam scores exists between students using Technique 1 and those using Technique 3.

Utilizing the Powerful Holm-Bonferroni Procedure

The **Holm Method**, often referred to as the Holm-Bonferroni Sequential Correction, is widely considered a superior alternative to the traditional Bonferroni correction. While it successfully maintains stringent control over the family-wise error rate (FWER), it achieves this using a sequential rejection process that dramatically enhances **statistical power**. This makes the Holm procedure the preferred choice for many researchers when conducting a fixed number of planned comparisons.

Due to its sequential nature and reduced conservatism, the Holm method is more likely to detect genuine differences between groups compared to the standard Bonferroni test. We implement this test using the same `pairwise.t.test()` function in R, adjusting the `p.adj` argument to `'holm'`:

```
#perform the Holm post-hoc method
pairwise.t.test(df$score, df$technique, p.adj='holm')
```

Pairwise comparisons using t tests with pooled SD

data: df$score and df$technique

tech1 tech2

tech2 0.206 -
tech3 0.048 0.384

P value adjustment method: holm

Examining the results, we find that the adjusted p-value for the difference between `tech1` and `tech3` remains 0.048. This confirms the significant difference identified by both the Tukey and Bonferroni methods. A key distinction highlighting the increased power of the **Holm Method** is the larger p-value for `tech2-tech1` found under Bonferroni (0.309) compared to Holm (0.206), although neither is significant.

Summary of Findings and Method Selection

The performance of the **ANOVA** omnibus test initially confirmed that a significant effect existed among the three study techniques. Subsequent application of four distinct **post-hoc comparisons** (Tukey, Scheffe, Bonferroni, and Holm) provided detailed insight into the specific group differences while controlling for the family-wise error rate.

A comparison of the results shows that three out of the four methods--Tukey HSD, Bonferroni, and Holm--concluded that the only **statistically significant difference** was found exclusively between Technique 1 and Technique 3. The Scheffe method, due to its inherent conservatism, failed to detect any significant pair, illustrating the potential loss of power when using highly conservative adjustments in balanced designs.

The choice of the appropriate post-hoc test relies heavily on the experimental design and the researcher's priority. For balanced designs comparing all possible pairs, the **Tukey HSD Method** is typically recommended for its optimal power. If the comparisons are planned and limited, the **Holm Method** offers a powerful and reliable alternative to the traditional Bonferroni correction.

Additional Resources for Post-Hoc Testing

The following tutorials provide additional information about **ANOVA** procedures and the theoretical application of various post-hoc tests in statistical analysis: