

How to perform a Bonferroni Correction in R

Authored by
stats writer

December 17, 2025

RECOMMENDED CITATION

stats writer (2025). *How to perform a Bonferroni Correction in R*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=107715>

When conducting statistical analysis, researchers often encounter scenarios where they must perform multiple hypothesis tests simultaneously. This situation, known as the multiple comparison problem, significantly increases the likelihood of finding a statistically significant result purely by chance--an unacceptable risk to scientific integrity. To counteract this inflation of false positives, statistical methods are employed to adjust the criteria for significance.

The Bonferroni correction stands out as one of the most straightforward and widely recognized techniques for managing this risk. It addresses the fundamental issue by adjusting the individual p-value required to maintain a consistent overall error rate across all tests performed. While highly conservative, its simplicity makes it a favorite in fields ranging from genetics to psychology.

In the powerful statistical environment of R, applying the Bonferroni correction is remarkably accessible, primarily via the built-in `p.adjust()` function. This function allows users to input a vector of raw p-values and specify the desired adjustment method, returning a new vector of adjusted p-values that account for the simultaneous nature of the tests. Throughout this guide, we will demonstrate how to correctly implement this procedure within R, specifically in the context of post-hoc analysis following an ANOVA.

The Role of Bonferroni in Post-Hoc Analysis

Before diving into the implementation of the Bonferroni adjustment, it is essential to understand the context in which it is most frequently applied: post-hoc analysis following an Analysis of Variance (ANOVA). A one-way ANOVA serves as an omnibus test, designed to determine whether there is any statistically significant difference among the means of three or more independent groups simultaneously. If the result is significant--meaning the overall F-test yields a p-value below the chosen significance level (e.g., $\alpha = 0.05$)--we conclude that the group means are not all equal.

However, the ANOVA itself provides no specific details regarding which pairs of groups differ from one another. It merely signals the presence of variance somewhere within the dataset. To isolate these specific differences, researchers must then conduct subsequent comparisons, typically in the form of multiple pairwise t-tests, comparing every possible combination of groups. If we have three groups (A, B, C), we need to run three separate t-tests: A vs. B, A vs. C, and B vs. C.

This process of performing multiple comparisons introduces the significant risk of inflating the Type I error rate. For instance, if we maintain an individual alpha level of 0.05 for each of three tests, the overall probability of making at least one false discovery (a Type I error) across the set of tests rapidly increases. This is precisely why we must implement a control mechanism, such as Bonferroni's correction, to manage the family-wise error rate (FWER).

Understanding the Family-Wise Error Rate (FWER)

The core problem that the Bonferroni correction addresses is the inflation of the FWER. The family-wise error rate is defined as the probability of making at least one Type I error (rejecting a true null hypothesis) when conducting a series, or family, of related hypothesis tests. If we perform k independent tests, each with an individual error rate of α , the FWER is calculated as $1 - (1 - \alpha)^k$. As k increases, this overall error rate quickly approaches 1, making our results highly unreliable.

For example, if we set $\alpha = 0.05$ and perform $k=5$ comparisons, the FWER is $1 - (1 - 0.05)^5 \approx 0.226$. This means there is a 22.6% chance of incorrectly declaring at least one comparison significant. In complex experiments involving dozens of comparisons, the FWER can become dangerously high, severely undermining the statistical validity of the findings.

The Bonferroni correction controls the FWER by proposing a simple adjustment to the individual comparison significance level. If a researcher wishes to maintain a desired FWER of α_{family} (typically 0.05) across k comparisons, the new, adjusted significance level for each individual test (α_{adjusted}) is calculated as $\alpha_{\text{adjusted}} = \alpha_{\text{family}} / k$. A raw p-value must then be smaller than this new, stricter threshold to be considered statistically significant. Alternatively, the Bonferroni method can adjust the raw p-values themselves by multiplying them by k , provided the resulting adjusted p-value does not exceed 1. This p-value adjustment is the method utilized by R's `p.adjust()` function.

Case Study: Comparing Study Techniques

To illustrate the application of Bonferroni's correction in R, let us consider a practical case study. Imagine an educational researcher studying the impact of various pedagogical interventions. Specifically, a teacher is interested in determining if three distinct studying techniques (labeled Tech 1, Tech 2, and Tech 3) result in significantly different performance metrics, measured by final exam scores among her students.

The experiment is designed rigorously: 30 students are randomly allocated to one of the three techniques, meaning 10 students are assigned to each group. After a standardized period of application (one week), all students take the same comprehensive exam. The analysis goal is twofold: first, determine if any overall difference exists using an ANOVA, and second, if a difference is found, use Bonferroni adjusted t-tests to pinpoint the exact group pairings that show statistically significant score disparities.

Step 1: Preparing and Structuring the Data in R

The first critical step in any statistical analysis using R is the correct creation and structuring of the

dataset. For this scenario, we must combine the continuous response variable (exam score) with the categorical predictor variable (studying technique) into a single data frame. The R code below demonstrates the creation of this structure, ensuring each student's score is correctly associated with their assigned technique.

We use the `data.frame()` function, employing `rep()` and `each` to efficiently generate the grouping variable. This results in 30 rows of data, 10 for each of the three levels of the independent variable.

#create data frame for 30 students across three groups

```
data <- 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))
```

#view first six rows of data frame to verify structure

```
head(data)
```

```
technique score
```

```
1 tech1 76
```

```
2 tech1 77
```

```
3 tech1 77
```

```
4 tech1 81
```

```
5 tech1 82
```

```
6 tech1 82
```

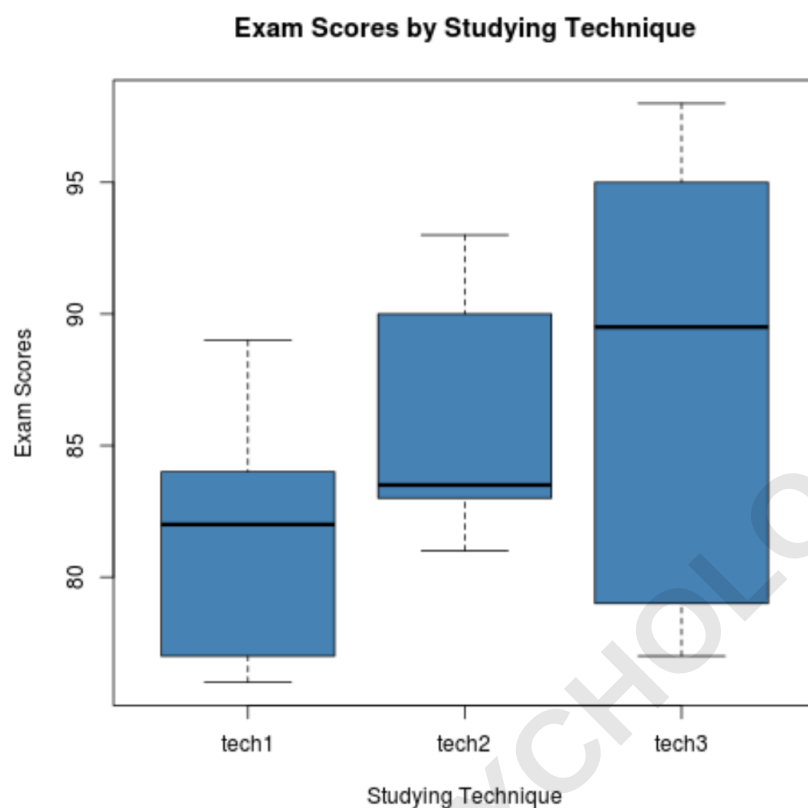
Step 2: Visualizing Group Differences

Before proceeding with formal inferential statistics, it is always recommended practice to visually inspect the data. Visualization provides initial insights into the distribution, central tendency, and potential variability within each group. For comparing quantitative data across categorical groups, a boxplot is an excellent choice, clearly displaying the median, quartiles, and range of scores for each study technique.

The R code below utilizes the base plotting functions to generate a comparative boxplot of the exam scores stratified by the study technique variable. This graphical representation helps us form initial hypotheses about where the differences might lie.

```
boxplot(score ~ technique,  
data = data,  
main = "Exam Scores by Studying Technique",  
xlab = "Studying Technique",
```

```
ylab = "Exam Scores",  
col = "steelblue",  
border = "black")
```



Upon reviewing the boxplot, we can visually observe that the mean score for technique 2 appears slightly higher than technique 1, and technique 3 shows a wider spread but also potentially a higher average score than technique 1. However, visual inspection is insufficient for making statistical claims; we must now proceed with the formal ANOVA test.

Step 3: Conducting the Omnibus One-Way ANOVA

The primary analysis begins with the one-way ANOVA. This test allows us to compare the means of the three groups simultaneously to determine if there is a statistically significant overall effect of the studying technique on the exam scores. In R, this is efficiently executed using the `aov()` function, which fits the ANOVA model.

The model structure used is `score ~ technique`, indicating that the exam score is the dependent variable predicted by the categorical factor `technique`. The output of the `summary()` function provides the ANOVA table, including the degrees of freedom (Df), sum of squares (Sum Sq), F-statistic (F value), and the critical overall p-value (Pr(>F)).

#fit the one-way ANOVA model

```
model <- aov(score ~ technique, data = data)
```

```
#view model output
```

```
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
```

Reviewing the ANOVA output, we observe that the overall p-value associated with the technique factor is **0.0476**. Since this value is less than the conventional significance level of $\alpha = 0.05$, we reject the null hypothesis. This finding strongly suggests that there is a statistically significant difference among the mean exam scores across the three studying techniques. Because the omnibus test is significant, we are justified in moving on to the post-hoc analysis to locate the specific differences.

Step 4: Implementing Bonferroni Adjusted Pairwise Tests in R

Since the ANOVA confirmed that differences exist, the next logical step is to perform pairwise t-tests to identify which specific pairs of means are significantly different. Crucially, we must incorporate the Bonferroni correction during this step to control the family-wise error rate, maintaining a reliable probability of Type I errors across the set of three comparisons (Tech 1 vs 2, Tech 1 vs 3, and Tech 2 vs 3).

In R, the most efficient way to perform these post-hoc tests with automatic FWER control is by using the powerful `pairwise.t.test()` function, which is part of the base `stats` package. This function is designed precisely for comparing multiple group means and offers various built-in methods for p-value adjustment, including the Bonferroni correction.

The syntax for the function is straightforward: `pairwise.t.test(x, g, p.adjust.method="bonferroni")`. Here, **x** represents the numeric response variable (our `data$score`), and **g** represents the grouping factor (our `data$technique`). Setting the `p.adjust.method` argument to `"bonferroni"` instructs R to calculate the adjusted p-values by multiplying the raw p-values by the number of comparisons ($k=3$).

Executing the function for our specific dataset provides the following results, displayed in a matrix format where comparisons are made between the row group and the column group:

```
#perform pairwise t-tests with Bonferroni's correction
```

```
pairwise.t.test(data$score, data$technique, p.adjust.method="bonferroni")
```

Pairwise comparisons using t tests with pooled SD

data: data\$score and data\$technique

```
tech1 tech2
```

```
tech2 0.309 -
```

```
tech3 0.048 1.000
```

P value adjustment method: bonferroni

Step 5: Interpreting the Bonferroni Adjusted P-Values

The resulting matrix from the `pairwise.t.test()` function provides the adjusted p-values for all possible comparisons between the study techniques. These adjusted values now directly control the FWER at the predetermined 0.05 level, allowing us to confidently assess which differences are statistically reliable.

We interpret the matrix by comparing the adjusted p-values to our original family-wise significance threshold ($\alpha_{\text{family}} = 0.05$). If an adjusted p-value is less than 0.05, we declare that pair of groups to be statistically significant. The output breaks down the results as follows:

The adjusted p-value comparing **Technique 1 and Technique 2** is **0.309**. Since this is much greater than 0.05, we find no statistically significant difference between these two studying methods.

The adjusted p-value comparing **Technique 1 and Technique 3** is **0.048**. This value is marginally less than 0.05, indicating a statistically significant difference between the exam scores obtained by students using Technique 1 versus Technique 3.

The adjusted p-value comparing **Technique 2 and Technique 3** is **1.000**. This result clearly shows no significant difference between the performance metrics of students using Technique 2 and Technique 3. Note that p-values are capped at 1.000 after adjustment, as they cannot exceed 1.

In conclusion, after controlling for the increased risk of Type I error using the Bonferroni correction, the analysis reveals that the only significant disparity in exam performance exists between **Technique 1 and Technique 3**. This means that while Technique 3 appears to lead to higher scores, it is only statistically superior when compared specifically against Technique 1, but not Technique 2.

A Note on the Conservatism of Bonferroni

While the Bonferroni correction is easy to implement and guarantees stringent control over the FWER, it is important to recognize its primary drawback: it is highly conservative. By strictly dividing the alpha level by the number of tests, it drastically reduces the statistical power of the individual tests, making it harder to detect true differences (increasing the risk of a Type II error, or false negative).

For research scenarios involving a very large number of comparisons, alternative methods that offer a better balance between Type I and Type II errors may be preferred. These include less conservative procedures such as the Holm-Bonferroni method (which is uniformly more powerful than the standard Bonferroni correction) or the Benjamini-Hochberg procedure, which controls the False Discovery Rate (FDR) rather than the FWER.

However, for a small number of planned comparisons, as is typical following a simple one-way ANOVA, the Bonferroni correction remains a robust and transparent choice, ensuring that any reported statistical significance is highly trustworthy and resistant to the inflation caused by multiple testing.

Further Resources on Statistical Testing in R

For researchers seeking to deepen their understanding of statistical comparisons and post-hoc methods in R, the following resources provide excellent foundations for expanding beyond the Bonferroni method:

[An Introduction to the One-Way ANOVA](#)

[Official R Documentation on Statistical Functions](#)

[Understanding Tukey's Honestly Significant Difference \(HSD\) Test](#)

[Introduction to Dunnett's Test for Comparing Against a Control](#)