

# Perform a Nested ANOVA in R (Step-by-Step)

Authored by  
**Mohammed loot**

November 5, 2025

## RECOMMENDED CITATION

Mohammed loot (2025). *Perform a Nested ANOVA in R (Step-by-Step)*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=10435>

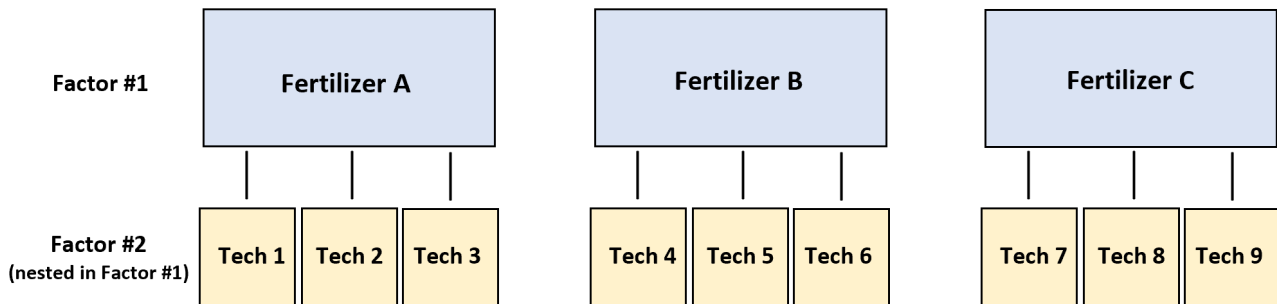
A [Nested ANOVA](#), frequently termed a Hierarchical [Analysis of Variance](#) (ANOVA), is a specialized and powerful statistical method essential for analyzing data derived from complex experimental designs. This technique is employed specifically when the structure of the experiment dictates that the levels of one experimental [factor](#) are entirely unique and confined within the levels of a higher-order, encompassing factor. Correctly identifying and modeling this hierarchical structure is paramount for accurately partitioning the total observed variability in the data and drawing valid statistical conclusions.

The necessity of the Nested ANOVA arises when standard two-way ANOVA assumptions--that all levels of one factor cross all levels of the other--are violated. If the interaction between factors is impossible or meaningless because the subunits are distinct across treatments, nesting becomes the appropriate modeling approach. Failure to account for this non-crossed structure can lead to inflated F-statistics and erroneous interpretations of treatment effects.

To illustrate this concept, consider a common practical scenario: an agricultural researcher is interested in evaluating the efficacy of three distinct types of fertilizer (labeled **A**, **B**, and **C**) on overall plant growth. To conduct this large-scale experiment, the researcher enlists the help of nine different field technicians, each operating independently.

The experimental protocol dictates a strict assignment: three distinct technicians are assigned exclusively to apply Fertilizer A, a separate group of three technicians applies Fertilizer B, and the final three technicians handle only Fertilizer C. Each technician is responsible for treating four individual plants. In this controlled setting, the primary continuous response variable is **plant growth**, and the two major categorical factors under investigation are **Fertilizer Type** and **Technician ID**.

The core statistical challenge here is that the **Technician** factor is definitively *nested* within the **Fertilizer** factor. Technician 1, who applies Fertilizer A, operates under completely different experimental conditions than Technician 4, who applies Fertilizer B. They are not comparable across all fertilizer types; rather, the technicians are specific subunits of the primary treatment group. This hierarchical relationship must be explicitly defined in the statistical model to avoid confounding the variability introduced by the fertilizer treatment with the variability potentially introduced by the individual technicians. The following schematic diagram visually represents this common nested structure, showing how the subunits are partitioned by the main factor:



This comprehensive tutorial provides a detailed, step-by-step methodology for executing, interpreting, and visualizing this specific type of nested ANOVA model using the highly versatile statistical programming language, [R](#).

## Step 1: Preparing and Structuring the Data in R

The foundational step in any statistical analysis is ensuring that the raw experimental data is correctly formatted and structured within the analytical environment. Before we can fit the nested model, we must organize our results into a suitable [data frame](#) within the [R](#) environment. Our resulting data frame, conventionally named `df` for simplicity, must contain three essential columns: `growth` (the continuous response measurement), `fertilizer` (the primary treatment factor), and `tech` (the lower-level, nested factor).

The code below is designed to meticulously generate sample data that mirrors the described experimental design. This ensures a total of 36 observations (12 measurements for each of the three fertilizer types). Crucially, the code correctly distributes the nine technicians across these groups, ensuring that technicians 1, 2, and 3 are exclusively associated with Fertilizer A, 4, 5, and 6 with Fertilizer B, and so on. This structure, featuring four repeated measurements per technician, mathematically establishes the necessary nesting relationship.

### #create data

```
df <- data.frame(growth=c(13, 16, 16, 12, 15, 16, 19, 16, 15, 15, 12, 15,
19, 19, 20, 22, 23, 18, 16, 18, 19, 20, 21, 21,
21, 23, 24, 22, 25, 20, 20, 22, 24, 22, 25, 26),
fertilizer=c(rep(c('A', 'B', 'C'), each=12)),
tech=c(rep(1:9, each=4)))
```

```
#view first six rows of data
```

```
head(df)
```

```
growth fertilizer tech
```

```
1 13 A 1
```

```
2 16 A 1
3 16 A 1
4 12 A 1
5 15 A 2
6 16 A 2
```

Carefully reviewing the output of the `head(df)` function is a critical verification step. The initial rows confirm that Technicians 1, 2, and 3 are indeed exclusively linked to the Fertilizer A category. This visual check validates that the data is prepared correctly, fulfilling the requirement for a nested design where the technician IDs are unique identifiers within the larger fertilizer groups. Proper data setup prevents model misidentification and ensures the statistical analysis proceeds accurately.

## Step 2: Implementing the Nested ANOVA Model using `aov()`

To execute the [Nested ANOVA](#) model in R, we rely on the standard `aov()` function, which is the cornerstone for fitting linear models in the base R environment. However, the syntax used within this function must be carefully constructed to explicitly communicate the hierarchical nature of the experimental factors to the statistical engine.

The key to specifying nesting in the R formula interface is the forward slash (`/`) operator. This operator is used to denote that the factor following the slash is nested within the factor preceding it. When the model is processed, R mathematically decomposes the total variance into the main effect (Factor A) and the nested effect (Factor B within Factor A).

The general structural syntax for fitting a nested model is elegantly simple, yet highly specific:

**`aov(response ~ factor A / factor B)`**

In the context of our plant growth experiment, the components translate directly:

**response:** This is `df$growth`, the continuous measurement of plant growth.

**factor A:** This is `df$fertilizer`, the main, higher-level categorical factor.

**factor B:** This is `df$tech`, the lower-level factor that is nested within the fertilizer groups.

A crucial technical detail in R is ensuring that the nested factor, `tech` (which contains numerical IDs 1 through 9), is correctly treated as a [categorical variable](#), rather than a continuous numeric variable. This is handled by wrapping it in the `factor()` function directly within the model call. Applying this precise syntax to our plant growth dataset yields the following, correctly specified model fit:

**#fit nested ANOVA**

```
nest <- aov(df$growth ~ df$fertilizer / factor(df$tech))
```

```
#view summary of nested ANOVA
```

```
summary(nest)
```

```
Df Sum Sq Mean Sq F value Pr(>F)
df$fertilizer 2 372.7 186.33 53.238 4.27e-10 ***
df$fertilizer:factor(df$tech) 6 31.8 5.31 1.516 0.211
Residuals 27 94.5 3.50
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Step 3: Interpreting the Analysis of Variance Table**

The output generated by `summary(nest)` presents the classical [ANOVA](#) table, which is the primary tool for assessing the statistical influence of the factors. Interpreting the results of a nested model requires careful attention to the rows, as they represent the variance sources partitioned by the model structure: the main factor (Fertilizer) and the nested term (Technician within Fertilizer).

We systematically analyze the results in the two main rows of the table to test our hypotheses:

**Main Factor (`df$fertilizer`):** This first row evaluates the overall impact of the primary experimental treatment--the type of fertilizer used. The computed F-value for this factor is exceptionally high (53.238), indicating a large amount of variability explained by the fertilizer types. Crucially, the corresponding [P-value](#) is extremely small (4.27e-10). Because this P-value is dramatically lower than the standard significance threshold ( $\alpha = 0.05$ ), we decisively reject the null hypothesis that all fertilizer means are equal. We conclude that the specific fertilizer type applied has a highly **statistically significant** effect on the resulting plant growth.

**Nested Factor (`df$fertilizer:factor(df$tech)`):** This row represents the variance introduced by the individual technicians, but only after accounting for the fact that they are uniquely assigned (nested) within the larger fertilizer groups. This term assesses whether differences in application technique, skill, or slight environmental variations introduced by the technicians significantly contribute to the overall variance. The F-value calculated here is 1.516, which yields a [P-value](#) of 0.211. Since 0.211 is considerably greater than the 0.05 alpha level, we fail to reject the null hypothesis for this nested factor. Therefore, we find no compelling evidence that the individual technician significantly impacts plant growth, once the powerful effect of the specific fertilizer type is already accounted for.

The overall interpretation strongly suggests that the overwhelming majority of the observed

variability in plant growth is attributable to the inherent differences between the fertilizer compositions (Factor A). The systematic differences or 'noise' introduced by the individual human applicators (Factor B) are statistically negligible relative to the treatment effect. If the experimental goal is maximizing yield, resources should be optimally focused on refining the composition of the fertilizer itself, as technician variability is not a major source of error.

## Step 4: Visual Confirmation and Exploratory Data Analysis with ggplot2

While the numerical precision of the [ANOVA](#) table provides definitive statistical evidence, visualizing the data remains an essential component of robust statistical practice. Visual inspection helps confirm the model's conclusions, highlights distribution characteristics, and provides a clearer, intuitive understanding of the nesting structure. [Boxplots](#), specifically, are exceptionally effective for comparing the center and spread of continuous data across multiple categorical groups.

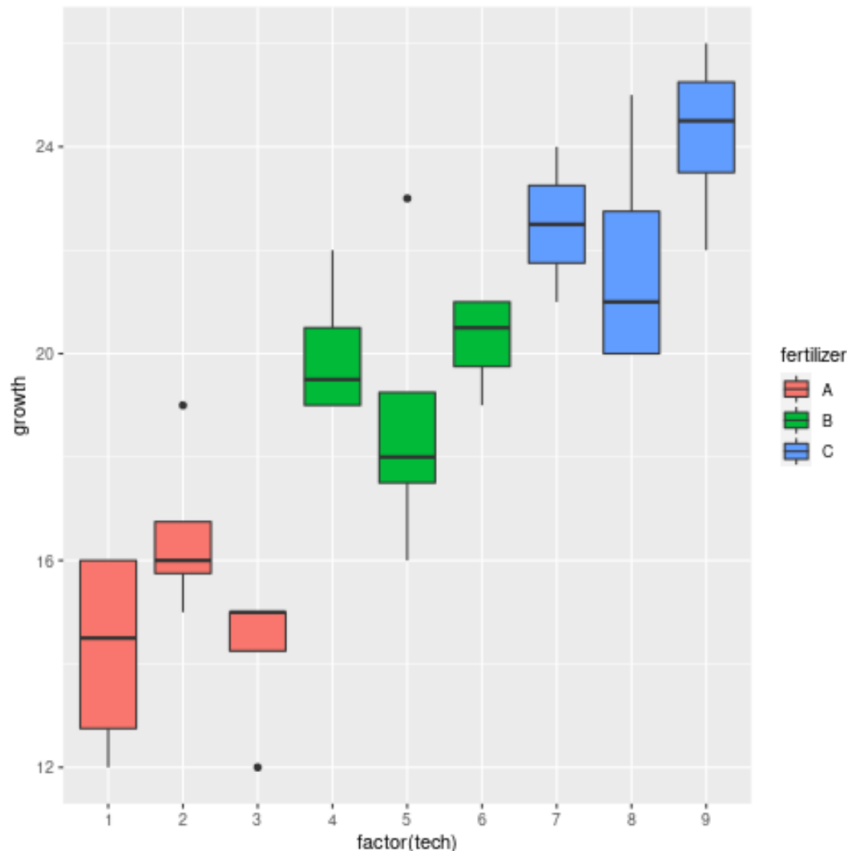
We leverage the capabilities of the powerful [ggplot2](#) package in [R](#) to construct an informative visualization. Our goal is to create a graphical representation that first groups the growth results by fertilizer (the primary factor) and then further separates the data by technician (the nested factor). This visual breakdown is vital for illustrating the concept of nesting, as it shows the subunits (technicians) operating only within their respective main treatment categories.

```
#load ggplot2 data visualization package
```

```
library(ggplot2)
```

```
#create boxplots to visualize plant growth  
ggplot(df, aes(x=factor(tech), y=growth, fill=fertilizer)) +  
geom_boxplot()
```

The resulting visualization effectively plots the technician ID on the x-axis, using color and grouping to delineate the fertilizer type they were assigned. The visual grouping clearly enforces the nesting: technicians 1, 2, and 3 are clustered together under one color, followed by 4, 5, and 6 under the next, and so on. This structure immediately makes the hierarchical nature of the data apparent.



Visually analyzing the boxplots reinforces the strong findings from the ANOVA table. We observe a clear and substantial upward progression in the median growth rates (indicated by the line inside the box) as we move from Fertilizer A groups toward Fertilizer C groups. This high inter-group variability confirms the **statistically significant** effect of the fertilizer. Conversely, when comparing the boxplots \*within\* any single fertilizer group (e.g., comparing technicians 1, 2, and 3), the distributions appear relatively similar in height and spread. This visual consistency confirms that the variability introduced by the individual technician is minimal compared to the variance introduced by the difference in the fertilizer treatments themselves.

## Step 5: Conclusion and Next Steps in Advanced Modeling

The successful application of the [Nested ANOVA](#) methodology allowed us to accurately decompose the total variance in plant growth into its constituent parts: the variance attributed to the major treatment factor (fertilizer) and the residual variance attributed to the nested subunit factor (technician). Our analysis provided compelling evidence of a strong, [statistically significant](#) effect of fertilizer type, while simultaneously confirming that the individual technician, nested within those fertilizer groups, exerted no significant influence on the outcome.

This approach to variance partitioning is invaluable across numerous scientific disciplines,

including biology, manufacturing quality control, and psychological research. It allows researchers to precisely separate legitimate treatment effects from systematic noise or differences inherent to nested experimental units--whether those units are people (technicians), physical samples (batches of reagents), or geographic locations (sub-plots within a larger zone). The ability to isolate these variance components ensures that research conclusions are robust and directly tied to the intended experimental manipulation.

For researchers seeking to build upon this foundational analysis, the next logical step involves checking the underlying assumptions of the ANOVA model. These assumptions include the normality of the residuals and the homogeneity of variances (homoscedasticity). Diagnostic plots, such as Residuals vs. Fitted values and Normal Q-Q plots, should be generated immediately following the model fitting process to ensure the validity of the F-tests and [P-values](#) reported in the summary table.

Additional Resources for Advanced Statistical Modeling and R Programming:

Official documentation and community support for the [R](#) statistical environment.

Detailed guides covering model selection, diagnostic checking, and assumption testing for general linear models.

Resources focused on applying mixed-effects models (which often extend nested designs to handle random effects).