

# How to Perform a One-Way ANOVA and Interpret the Results by Hand

Authored by  
**stats writer**

March 13, 2026

## RECOMMENDED CITATION

stats writer (2026). *How to Perform a One-Way ANOVA and Interpret the Results by Hand*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=135550>

## Understanding the Fundamentals of Analysis of Variance

The **Analysis of Variance**, commonly referred to as **ANOVA**, represents one of the most fundamental statistical techniques used in experimental research to determine if there are significant differences between the means of multiple groups. While a standard t-test is suitable for comparing two groups, the **One-way ANOVA** is specifically designed to handle three or more independent groups simultaneously. By utilizing this method, researchers can avoid the problem of **Type I error** inflation, which occurs when performing multiple pairwise comparisons. Instead of testing each pair individually, ANOVA provides a single global test to see if at least one group mean differs significantly from the others.

To perform a **one-way ANOVA**, we must satisfy several underlying assumptions to ensure the validity of our results. First, the **independence of observations** is paramount, meaning that the data points in one group should not influence those in another. Second, the data within each group should follow a **normal distribution**, particularly when sample sizes are small. Finally, we assume **homogeneity of variance**, implying that the variance among the groups is approximately equal. When these conditions are met, the **ANOVA** procedure offers a robust framework for identifying patterns within complex datasets.

The core logic of **ANOVA** involves partitioning the total variability found in a dataset into two distinct components: the variability between the group means and the variability within the groups themselves. If the variability between the groups is significantly larger than the variability within the groups, we can conclude that the group identity has a measurable effect on the outcome variable. This mathematical decomposition allows us to derive the **F-statistic**, a ratio that serves as the primary indicator of statistical significance. Understanding how to perform these calculations by hand provides deep insight into how **statistical significance** is actually determined.

In the following sections, we will explore a detailed example involving academic performance to illustrate the step-by-step process of conducting a **one-way ANOVA**. By manually calculating the **sum of squares**, **degrees of freedom**, and the final **F-ratio**, you will gain a comprehensive understanding of the mechanics behind the software outputs. This foundational knowledge is essential for any student or researcher looking to master **quantitative research** methods and interpret data with high precision and confidence.

## Establishing the Research Scenario and Hypotheses

Before diving into the mathematical calculations, we must clearly define our experimental setup. Imagine a scenario where an educational researcher wants to evaluate the efficacy of three distinct **exam prep programs**. The goal is to determine if one program is superior to the others or if all three yield essentially the same results. To investigate this, the researcher recruits 30 students and

randomly assigns 10 students to each of the three programs. Random assignment is critical here, as it helps mitigate **confounding variables** that might otherwise bias the final test scores.

After three weeks of rigorous preparation using their assigned programs, all 30 students take the same standardized exam. The resulting scores represent our **dependent variable**, while the type of prep program serves as our **independent variable**. In this context, the **null hypothesis** ( $H_0$ ) posits that there is no difference between the **population means** of the three groups. Conversely, the **alternative hypothesis** ( $H_a$ ) suggests that at least one of the group means is different from the others. We will use the **ANOVA** procedure to decide whether to reject or fail to reject this null hypothesis.

The visual representation of our data is the first step in the analysis. By organizing the scores into a clear table, we can begin to see potential trends. However, visual inspection alone is insufficient for scientific proof. We require a rigorous mathematical framework to account for the **sampling error** and natural variation inherent in any human-based study. The raw scores for our three groups are presented in the following image, providing the basis for all subsequent calculations in this tutorial.

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81

As we move forward, keep in mind that the total **sample size** ( $n$ ) is 30, and the number of groups ( $k$ ) is 3. These constants will be used repeatedly throughout the **ANOVA** steps. Our objective is to determine if the observed differences in exam scores are likely due to the prep programs themselves or simply the result of random chance. Through the meticulous application of the **ANOVA** formula, we will transform these raw numbers into a meaningful statistical conclusion that can inform educational policy and practice.

## Step 1: Calculating Group Means and the Grand Mean

The first practical step in performing a **one-way ANOVA** by hand is to calculate the **arithmetic**

**mean** for each of the individual groups, as well as the overall mean of the entire dataset. The group mean provides a central value that represents the typical performance within a specific condition. To find the mean for Group 1, we sum all ten scores and divide by the number of observations in that group. We repeat this process for Group 2 and Group 3. These **sample means** serve as our primary point of comparison when evaluating the effectiveness of the prep programs.

Once the individual group means are established, we calculate the **overall mean**, also known as the **grand mean**. The grand mean can be calculated in two ways: by averaging all 30 individual scores or by taking the average of the three group means (provided the group sizes are equal). The grand mean acts as a baseline, representing the average performance of all participants regardless of which prep program they used. The deviations of the individual group means from this grand mean are what we use to measure the variability between the groups.

In our specific example, the calculations yield distinct values for each prep program. These values allow us to see at a glance how the groups performed relative to one another. For instance, if one group mean is substantially higher than the grand mean, it suggests that the corresponding prep program might be more effective. However, we must quantify this observation through the **sum of squares** to determine if the difference is **statistically significant**. The calculated means for our dataset are displayed in the image below, which serves as a reference for the next phase of the **ANOVA** process.

	<b>Group 1</b>	<b>Group 2</b>	<b>Group 3</b>
	85	91	79
	86	92	78
	88	93	88
	75	85	94
	78	87	92
	94	84	85
	98	82	83
	79	88	85
	71	95	82
	80	96	81
<b>Group Means</b>	<b>83.4</b>	<b>89.3</b>	<b>84.7</b>
<b>Overall Mean</b>	<b>85.8</b>		

As shown in the data, the means for the three groups are 83.4, 89.3, and 84.7, respectively. The **grand mean** ( $\bar{X}$ ) is calculated to be 85.8. These figures are the building blocks for our **variability** analysis. By understanding where the center of each group lies, we can begin to measure how much the "treatment" (the prep program) moved the students away from the overall average. This sets the stage for calculating the **Sum of Squares Regression (SSR)**, which focuses on these

specific differences.

## Step 2: Quantifying Variability Between Groups (SSR)

After determining the means, we must calculate the **Sum of Squares** Between Groups, often labeled as **SSR** (Sum of Squares for Regression) or SSB. This metric quantifies the portion of the total variation that can be attributed to the differences between the group means. In simpler terms, SSR measures how much the average score of each prep program deviates from the overall average of all students. If the prep programs have a strong effect, we expect the group means to be spread far apart from the grand mean, resulting in a larger SSR.

The mathematical formula for SSR is  $n\sum(X_j - \bar{X})^2$ . Here, "n" represents the number of observations in each group, " $X_j$ " is the mean of a specific group, and " $\bar{X}$ " is the grand mean. For each group, we subtract the grand mean from the group mean, square the result to ensure all values are positive, and then multiply by the group size. Finally, we sum these values together. This weighting by "n" is essential because it accounts for the amount of data supporting each group's mean. A large SSR relative to the error suggests that the **independent variable** is exerting a significant influence on the **dependent variable**.

Applying this to our exam prep data, we perform the following calculation:  $10(83.4 - 85.8)^2 + 10(89.3 - 85.8)^2 + 10(84.7 - 85.8)^2$ . Breaking this down, we see that the differences are squared to penalize larger deviations more heavily. After completing the arithmetic, we find that our **SSR equals 192.2**. This single number represents the total "signal" or the explained variance in our study. However, to determine if this signal is meaningful, we must compare it to the "noise" or unexplained variance within the groups, which we will calculate in the next step.

Understanding **SSR** is vital because it directly reflects our research hypothesis. If all prep programs were identical in their effectiveness, we would expect the group means to be very close to the grand mean, leading to an SSR near zero. The fact that our SSR is 192.2 indicates that there is indeed some level of variation between the groups. The remaining question is whether this variation is large enough to be considered a real effect or if it is just a byproduct of the natural **variance** found within any sample of students.

## Step 3: Calculating the Within-Group Variability (SSE)

The next component of our **one-way ANOVA** is the **Sum of Squares Error (SSE)**, also known as the Sum of Squares Within Groups (SSW). While SSR measures the differences between groups, SSE measures the variation within each individual group. This variation represents the differences among students who used the *same* prep program. Since these students received the exact same treatment, any differences in their scores must be due to individual background, study habits, or random factors. In **statistical modeling**, SSE is often viewed as the "noise" that can obscure the

"signal" of the treatment effect.

The formula for SSE is  $\sum(\mathbf{X}_{ij} - \mathbf{X}_j)^2$ , where "X<sub>ij</sub>" represents each individual observation and "X<sub>j</sub>" is the mean of the group to which that observation belongs. To calculate this by hand, we take every single score in the dataset, subtract its corresponding group mean, square that difference, and then sum all 30 of those squared values. This process is more time-consuming than calculating SSR because it involves every data point, but it is necessary for accurately assessing the **standard deviation** within the groups. A lower SSE indicates that the members of a group performed very similarly to one another.

For our study, the SSE calculation is divided into three parts. For Group 1, we subtract the mean of 83.4 from each of the 10 scores, square them, and sum them to get 640.4. For Group 2, we subtract 89.3 from each score, resulting in a sum of 208.1. Finally, for Group 3, we subtract 84.7 from each score, giving us 252.1. Adding these three components together (640.4 + 208.1 + 252.1) results in a total **SSE of 1100.6**. This value quantifies the inherent variability among the students that the prep programs cannot explain.

By calculating **SSE**, we provide a context for our **SSR**. In our case, an SSE of 1100.6 is considerably larger than our SSR of 192.2. This suggests that there is quite a bit of "noise" in our data--individual students vary significantly in their test-taking abilities regardless of the program they used. This high internal variation will make it more difficult to prove that the differences between the programs are statistically significant. The relationship between these two sums of squares is what ultimately determines the outcome of the **F-test**.

#### Step 4: Determining the Total Sum of Squares (SST)

The **Total Sum of Squares (SST)** represents the overall variation in the entire dataset. It is the sum of the squared differences between each individual observation and the grand mean. One of the most elegant properties of **ANOVA** is the additive nature of these variations: the Total Sum of Squares is exactly equal to the sum of the Between-Group variation (SSR) and the Within-Group variation (SSE). This relationship, **SST = SSR + SSE**, serves as a crucial cross-check for our manual calculations. It confirms that we have accounted for every bit of variation present in the 30 exam scores.

In our example, calculating the SST is straightforward since we have already derived the values for SSR and SSE. By adding 192.2 (the variation explained by the prep programs) and 1100.6 (the variation due to individual differences), we arrive at an **SST of 1292.8**. This number tells us the absolute amount of "squared deviation" present in our exam scores relative to the grand mean of 85.8. It provides a holistic view of the dataset's spread before we begin the formal process of hypothesis testing.

Beyond being a simple sum, **SST** is important because it allows us to calculate the **coefficient of determination** (R-squared) if we choose to. R-squared is the ratio of SSR to SST, indicating the proportion of the total variation that is explained by our independent variable. In our case, 192.2 divided by 1292.8 would tell us what percentage of the exam score differences can actually be attributed to the prep programs. While not always required in a basic **ANOVA** table, it offers valuable context regarding the **effect size** of the study.

At this stage, we have successfully partitioned the data's variability. We have quantified the "signal" (SSR) and the "noise" (SSE), and verified them against the "total" (SST). However, these sums of squares cannot be compared directly because they are based on different numbers of observations and groups. To make them comparable, we must adjust them by their respective **degrees of freedom** to calculate the Mean Squares. This transition from "Sum of Squares" to "Mean Squares" is what allows us to finally compute the F-statistic.

### Step 5: Organizing Data into the ANOVA Summary Table

The **ANOVA table** is a standardized way of presenting the results of our calculations. It organizes the **Source of Variation**, **Sum of Squares (SS)**, **Degrees of Freedom (df)**, **Mean Squares (MS)**, and the final **F-statistic** into a clean, readable format. This table is what you will typically see in **academic journals** and statistical software outputs. By filling in this table, we synthesize all our previous work into a single diagnostic tool for our hypothesis test.

The **Degrees of Freedom (df)** are calculated based on the number of groups ( $k$ ) and the total number of observations ( $n$ ). For the treatment (SSR), the df is  $k - 1$ , which in our case is  $3 - 1 = 2$ . For the error (SSE), the df is  $n - k$ , or  $30 - 3 = 27$ . The total df is  $n - 1$ , which is 29. These values are essential because they adjust the Sum of Squares for the size of the study, resulting in the **Mean Squares (MS)**. We calculate MS by dividing the SS by its corresponding df ( $MS = SS / df$ ). The MS Treatment (96.1) and MS Error (40.8) represent the average variance between groups and within groups, respectively.

Source	Sum of Squares (SS)	df	Mean Squares (MS)	F
Treatment	192.2	2	96.1	2.358
Error	1100.6	27	40.8	
Total	1292.8	29		

The final step in completing the table is calculating the **F-statistic**. This is done by dividing the **MS Treatment** by the **MS Error** ( $96.1 / 40.8$ ). Our resulting **F-value is 2.358**. This ratio tells us how much larger the between-group variance is compared to the within-group variance. If the null hypothesis were true, we would expect the F-value to be close to 1. An F-value significantly greater

than 1 suggests that the differences between the group means are larger than what would be expected by random chance alone.

With the **ANOVA table** complete, we now have a single number--2.358--that summarizes the entire experiment. However, this number does not tell us whether the result is significant on its own. To make that determination, we must compare our calculated F-value to a **critical value** from the **F-distribution**. This comparison will allow us to conclude whether the observed differences in exam prep programs are meaningful or if they should be dismissed as statistical noise.

## Step 6: Interpreting the F-Statistic and Reaching a Conclusion

To interpret our **F-statistic** of 2.358, we must refer to the **F-distribution table**. This table provides the **critical value** that the F-statistic must exceed for the results to be considered statistically significant at a chosen **significance level** (alpha), typically set at 0.05. The critical value depends on two separate degrees of freedom: the numerator degrees of freedom ( $df_1 = 2$ ) and the denominator degrees of freedom ( $df_2 = 27$ ). These values define the specific shape of the F-distribution curve for our study.

Consulting an F-table with **alpha = 0.05**, **df1 = 2**, and **df2 = 27**, we find that the **F critical value** is **3.3541**. This value represents the threshold of significance. If our calculated F-statistic were higher than 3.3541, it would fall into the "rejection region," allowing us to reject the null hypothesis and conclude that the prep programs made a significant difference. However, in our analysis, the calculated F-statistic of 2.358 is clearly *less than* the critical value of 3.3541.

Because our F-statistic does not exceed the critical value, we **fail to reject the null hypothesis**. In practical terms, this means that while there were differences in the average exam scores between the three programs, those differences were not large enough to rule out random chance. The variability within the groups was too high relative to the differences between the group means. Consequently, we do not have sufficient evidence to claim that any one exam prep program is more effective than the others based on this specific study and **sample**.

This conclusion is an important part of the scientific process. Failing to find a significant difference is just as valuable as finding one, as it prevents us from making false claims about the effectiveness of a treatment. It may suggest that the researcher needs a larger sample size to detect a smaller **effect size**, or that the prep programs are indeed very similar in their impact. By performing this **ANOVA** by hand, we have followed the rigorous logic of statistical inference to reach a reliable, evidence-based conclusion.

## Conclusion and Final Thoughts on Manual ANOVA

Performing a **one-way ANOVA** by hand is a rigorous but rewarding exercise that demystifies the

process of **hypothesis testing**. By systematically calculating the group means, partitioning the sum of squares, and deriving the F-statistic, we gain a clear view of how variation is analyzed in the social and natural sciences. This manual approach ensures that we don't just treat **statistical software** as a "black box," but rather understand the mathematical principles that govern our conclusions.

In our exam prep example, we learned that even when means appear different on the surface, they may not be **statistically significant** when the internal "noise" of the data is taken into account. This highlights the importance of the **SSE** component and the role of the **F-ratio** in balancing explained vs. unexplained variance. For future studies, researchers might use these insights to refine their experimental designs, perhaps by controlling for additional variables to reduce the **within-group variance** and increase the sensitivity of the test.

While manual calculations are excellent for learning, modern researchers often rely on automated tools for larger datasets and complex analyses. If you are working with more than three groups or a very large number of observations, using a dedicated **one-way ANOVA calculator** can save time and reduce the risk of arithmetic errors. These tools follow the exact same steps we have outlined here, providing you with the **p-value** and F-statistic instantaneously. Whether you calculate by hand or use software, the fundamental goal remains the same: to turn raw data into actionable knowledge through the power of **statistics**.