# Analysis of Variance

**B M King,** Clemson University, Clemson, SC, USA



## Glossary

**Alpha** – The probability of a type I error (rejection of a true null hypothesis).
**Analysis of variance (ANOVA)** – A statistical test used to compare the means of three or more populations.
**Factor** – An independent (treatment) variable.
***F* ratio** – $s^2_{bet}/s^2_w$; between-groups variance estimate divided by the within-groups variance estimate.
**Grand mean** – The mean of all scores in all the treatment conditions.
**Interaction effect** – In analysis of variance, the joint effect of two or more independent variables that cannot be predicted from the main effects.
**Level(s)** – Different values of an independent variable (factor).
**Main effect** – In analysis of variance, the mean differences among the levels of one factor averaged across the levels of the other factor.
**Omega-squared** – $\hat{\omega}^2_{bet}$; a measure of effect size for analysis of variance.
**Orthogonal comparisons** – Comparisons that are unique and do not overlap.
**Planned comparisons** – Statistical tests used in place of ANOVA when the researcher knows in advance which comparisons are of interest.
***Post hoc* comparisons** – Statistical tests performed after obtaining a significant value of *F* that indicates which means are significantly different.
**Sum of squares (*SS*)** – Sum of the squared deviations from the mean.
**Type I error** – Rejection of a true null hypothesis.

Imagine that you are superintendent of a school district and you wish to compare the mathematics competency of students enrolled in seven high schools. You randomly select 30 seniors from each school and give them a mathematics achievement test. Your null hypothesis is that the mean mathematics achievement scores in the seven populations (the seven high schools) are equal, that is, there is no difference in scores among the schools. How should you analyze the results? If you had decided to test for differences between all pairs of means with use of a *t* test, you would have to conduct 21 tests. Suppose that you had decided in advance to set the probability of a type I error (rejection of a true null hypothesis) at 0.05. This means that when testing for a difference between two population means, in 20 replications of the same study (drawing independent random samples of the same size each time) you would expect the results of one of the *t* tests to be statistically significant even if there is no difference between the population means. Although for any one test, alpha (probability of a type I error) is 0.05, with 21 independent tests the probability of making at least one type I error is $p = 1 - (0.95)^{21} = 1 - 0.34 = 0.66$. Thus, if you found a statistically significant difference for one or two of your 21 tests, there is a high probability that the result(s) was due to a type I error rather than to a true difference between groups.

You need a test that allows you to make multiple comparisons among means while holding alpha steady at a preselected level. To this end, Ronald Fisher (1890–1962) developed analysis of variance (ANOVA) in the 1920s. An interesting feature of ANOVA is that although it is a test for differences among population means, it can be conducted without calculating the value of any of the sample means.

## The Logic of ANOVA

When testing for differences among population means, variation in scores can be due to two things: chance and treatment effect. As the scores in any one sample are drawn randomly from the same population, the scores can differ from one another, and from the sample mean, only due to chance. Scores in different samples, as well as sample means, can differ from one another due to chance, but also because the experimental treatment has an effect. If there is no treatment effect, scores in different groups will differ only because of chance. Thus, when the null hypothesis (there is no treatment effect) is true, the variation of scores around the sample means within each sample $(X - \overline{X})$ and the variation of sample means from the grand mean $(\overline{X} - \overline{\overline{X}}$; the grand mean, $\overline{\overline{X}}$, is the mean of all scores) will both be a reflection of variation in scores due to chance factors.

For any study of two or more samples, the amount by which a score differs from the grand mean $(X - \overline{\overline{X}})$ can be partitioned into two parts: the amount by which the score differs from its own sample mean $(X - \overline{X})$ and the amount by which the sample mean differs from the grand

mean $(\overline{X} - \overline{\overline{X}})$. The partition of $X - \overline{\overline{X}}$ is represented by the following equation:

$$X - \overline{\overline{X}} = (X - \overline{X}) + (\overline{X} - \overline{\overline{X}}) \quad [1]$$

As the name suggests, in ANOVA we are going to work with variance estimates. Variance estimates are calculated by dividing the sum of the squared deviations from the mean (often called sum of squares, and abbreviated *SS*) by the degrees of freedom (*df*). We can calculate the sum of squares total ($SS_{\text{total}}$) for two or more samples by calculating the deviation from the grand mean for every score, squaring each deviation score, and then summing ($\sum$) all the squared deviations from the grand mean:

$$SS_{\text{total}} = \sum^{\text{all scores}} (X - \overline{\overline{X}})^2$$

Look again at eqn [1]. You can see that we could also calculate $SS_{\text{total}}$ by squaring $(X - \overline{X}) + (\overline{X} - \overline{\overline{X}})$:

$$\sum^{\text{all scores}} (X - \overline{\overline{X}})^2 = \sum^{\text{all scores}} (X - \overline{X})^2 + \sum^{k} n\,(\overline{X} - \overline{\overline{X}})^2 \quad [2]$$

where:

$n$ = the number of scores in a particular group
$k$ = the number of samples

When we square $(a + b)$ there is a middle term, here $2(\overline{X} - \overline{\overline{X}}) \sum (X - \overline{X})$, but in this case it is always zero. What has been accomplished is the partitioning of the total sum of squares into two parts: (1) $\sum^{\text{all scores}} (X - \overline{X})^2$, called $SS_{\text{within}}$ (or $SS_{\text{error}}$), a reflection only of chance factors when drawing samples, and (2) $\sum^{k} n\,(\overline{X} - \overline{\overline{X}})^2$, called $SS_{\text{between}}$, a reflection of both chance factors and treatment effect (when the population means are different). Thus,

$$SS_{\text{total}} = SS_{\text{within}} + SS_{\text{between}}$$

We can now divide the separate sums of squares by their respective degrees of freedom to obtain variance estimates. The degrees of freedom for $SS_{\text{total}}$ is the total number of scores minus one: $n_{\text{total}} - 1$. For $SS_{\text{within}}$ (abbreviated $SS_{\text{w}}$), the degrees of freedom is calculated by adding $n$-1 for each sample: $n_{\text{total}} - k$. For $SS_{\text{between}}$ (abbreviated $SS_{\text{bet}}$), degrees of freedom is the number of samples minus one: $k - 1$.

The end result is two variance estimates. One is the within-groups variance estimate, $s^2_{\text{w}}$ $(= SS_{\text{w}}/df_{\text{w}})$, which is an estimate of the variance due to chance. It is sometimes called mean square within or mean square error. The other is the between-groups variance estimate, $s^2_{\text{bet}}$ $(= SS_{\text{bet}}/df_{\text{bet}})$, which is an estimate of the variance due to chance plus any treatment effect. When there is no treatment effect, both are estimates of the same thing. Thus, when the null hypothesis is true (i.e., there is no treatment effect), the ratio $s^2_{\text{bet}}/s^2_{\text{w}}$, called the *F* ratio, should be about 1.0. When the null hypothesis is false, $s^2_{\text{bet}}$ will exceed $s^2_{\text{w}}$, and the greater the treatment effect, the more is the amount by which $s^2_{\text{bet}}$ will exceed $s^2_{\text{w}}$. Due to sampling variation, we expect the value of the *F* ratio to be different with every replication of a study (employing new randomly drawn samples of the same size). However, if we set alpha = 0.05 and obtain a value of *F* that would have occurred by chance less than 5% of the time (if the null hypothesis were true) we reject the null hypothesis and conclude that there is a treatment effect (i.e., the treatment-effect reliably exceeds effects due to chance).

## Computation of *F*: An Example

Let us look at an example. Suppose that we wish to compare two new methods for teaching mathematics with the standard method. We randomly draw a sample of ten subjects for each method and use each method for 1 year, at the end of which we administer a mathematics achievement test to each subject. The hypothetical results are as follows:

| Standard method ($X$) | Method 2 ($Y$) | Method 3 ($Z$) |
|---|---|---|
| 68 | 85 | 62 |
| 70 | 90 | 76 |
| 75 | 80 | 74 |
| 65 | 78 | 70 |
| 55 | 82 | 58 |
| 80 | 86 | 78 |
| 72 | 92 | 74 |
| 78 | 76 | 80 |
| 60 | 94 | 65 |
| 74 | 84 | 75 |
| $\sum X = 697$ | $\sum Y = 847$ | $\sum Z = 712$ |
| $\overline{X} = 69.7$ | $\overline{Y} = 84.7$ | $\overline{Z} = 71.2$ |
| $\overline{\overline{X}} = 2256/30 = 75.2$ | | |

We now calculate $SS_{\text{total}}$, $SS_{\text{w}}$, and $SS_{\text{bet}}$:

$$\begin{aligned} SS_{\text{total}} &= (68 - 75.2)^2 + (70 - 75.2)^2 + \cdots (65 - 75.2)^2 \\ &\quad + (75 - 75.2)^2 = 2722.8 \\ SS_{\text{w}} &= (68 - 69.7)^2 + (85 - 84.7)^2 + (62 - 71.2)^2 \\ &\quad + (70 - 69.7)^2 + (90 - 84.7)^2 + (76 - 71.2)^2 \\ &\quad \vdots \quad \vdots \quad \vdots \\ &\quad + (74 - 69.7)^2 + (84 - 84.7)^2 + (75 - 71.2)^2 \\ &= 1357.8 \\ SS_{\text{bet}} &= 10(69.7 - 75.2)^2 \\ &\quad + 10(84.7 - 75.2)^2 \\ &\quad + 10(71.2 - 75.2)^2 \\ &= 1365 \end{aligned}$$

There are formulas to calculate $SS_{\text{total}}$, $SS_{\text{bet}}$, and $SS_{\text{w}}$ that are easier to use than using deviations from means,

but they do not provide you with a basic understanding of ANOVA (see King and Minium, 2008). (It is these formulas that allow you to compute $F$ without ever having calculated a single mean.) It is customary to place the results in a summary table, often called an ANOVA table:

| Source | $SS$ | $df$ | $s^2$ | $F$ |
|---|---|---|---|---|
| Between groups | 1365 | 2 | 682.5 | 13.571 |
| Within groups | 1357.8 | 27 | 50.29 | |
| Total | 2722.8 | 29 | | |

Be sure to place between groups first, above within groups, to set up the proper ratio. As a check on your work, $SS_{bet}$ plus $SS_w$ should equal $SS_{total}$, and $df_{bet}$ plus $df_w$ should equal $df_{total}$ ($s^2_w$ plus $s^2_{bet}$ do not equal $s^2_{total}$).

In order to determine if our obtained value of $F$ is statistically significant, we must use a table for the $F$ distribution, found in any introductory statistics textbook. The table gives critical values of $F$. When our obtained value of $F$ is equal to or greater than the critical value, it indicates that our result is statistically significant (i.e., has a low probability of occurring by chance). The $F$ distribution varies as a function of $df_{bet}$ and $df_w$. Critical values of $F$ are found at the intersection of $df$ for the numerator (2 in our example) and $df$ for the denominator (27 in our example), and in this case the critical values are 3.35 for alpha = 0.05 and 5.49 for alpha = 0.01. Thus, in our example, the probability of a type I error was less than 0.01, that is, the result is statistically significant. In addition to your $F$ statistic and $p$-value, you should also provide a measure of effect size when reporting results. One commonly used measure for independent-groups ANOVA is omega-squared ($\hat{\omega}^2_{bet}$) (see King and Minium, 2008). It gives us a population-based, rather than a sample-based, estimate of how much of the proportion of the variance in the dependent variable was attributed to the different levels of treatment.

## Interpreting a Significant *F* Value

Independent-groups ANOVA can be used with two samples, in which case $F$ is the square of the $t$-statistic that compares the two sample means. A statistically significant result indicates that one population mean is either less than or greater than the other. What does it mean when we obtain a statistically significant value of $F$ for three or more samples? In this case it tells us only that there is a difference among the populations. It does not tell us the manner in which they differ. For three groups, all three population means could be different from one another, or one could be greater than the other two, etc. To determine which means are significantly different from others, we normally use *post hoc* (*a posteriori*) comparisons. Some of the most commonly used tests are Duncan's multiple-range test, the Newman–Keuls test, Tukey's HSD test, and the Scheffé test. Duncan's test is the least conservative with regard to type I error and the Scheffé test is the most conservative. An explanation of these tests is beyond the scope of this article, but most textbooks will provide a full explanation of one or more of them. However, before you can use any of them you must first have obtained a significant value of $F$. In our example, all four *post hoc* tests would reveal that teaching method 2 is superior to the other two methods, which did not significantly differ from one another.

There are some underlying assumptions associated with the use of ANOVA. The first is that the populations from which the samples are drawn are normally distributed. Moderate departure from the normal bell-shaped curve does not greatly affect the outcome, especially with large-sized samples (Glass *et al.*, 1972). However, results are much less accurate when populations of scores are very skewed or multimodal (Tomarken and Serlin, 1986), which is frequently the case in the behavioral sciences (Micceri, 1989). In this case, you should consider using the Kruskal–Wallis test, an assumption-freer (nonparametric) test for the independent-groups design (see King and Minium, 2008). This is especially true when using small samples. A second assumption is that of homogeneity of variance, that is, the variances in the populations from which samples are drawn are the same. However, this is a major problem only when variances differ considerably, and is less of a problem if you use samples that are of the same size (Milligan *et al.*, 1987; Tomarken and Serlin, 1986).

## The Repeated-Measures Design

ANOVA can also be used with the repeated-measures design (e.g., testing the same subjects under two or more conditions). The major difference from the independent-groups design is that $SS_w$ is partitioned into two parts: (1) $SS_{subjects}$, which is a reflection of the variability in scores resulting from individual differences, and (2) $SS_{residual}$, a reflection of the variability in scores due to chance. Although three variance estimates can be calculated ($s^2_{subjects}$, $s^2_{bet}$, and $s^2_{resid}$), normally we are interested only in the $F$ ratio $s^2_{bet}/s^2_{resid}$ to determine if there is a statistically significant difference among groups. (We are usually not interested in whether or not individuals differ.)

## Two-Way ANOVA

In the previous example, we examined the effects of three different teaching methods on mathematics achievement scores using an independent-groups design. There was only one treatment variable – teaching method. However, ANOVA allows us to study two or more treatment variables (called factors) simultaneously. Suppose, for example, that

we were also interested in whether the amount of time devoted to teaching mathematics affected achievement scores. Rather than conducting two separate studies, we can study both factors (teaching method and time spent teaching) simultaneously using a two-factor ANOVA design. If we study all three teaching methods (referred to as levels of the factor teaching method) and two different durations of time spent teaching (two levels; e.g., 45 min per day vs. 90 min), we have a 3 × 2 design. Let us suppose that we select eight students randomly for each of the six independent conditions (standard teaching method for 45 min per day, standard teaching method for 90 min per day, etc.) and after 1 year we obtain the following mathematics achievement scores:

| | *Teaching method* | | | |
|---|---|---|---|---|
| *Teaching time per day* | *Standard method* | *Method 2* | *Method 3* | |
| | 60 | 73 | 62 | |
| | 80 | 87 | 74 | |
| | 68 | 76 | 63 | |
| | 72 | 84 | 73 | |
| 45 min | 64 | 83 | 64 | $\overline{X}_{45} = 72.67$ |
| | 76 | 77 | 72 | |
| | 69 | 81 | 67 | |
| | 71 | 79 | 69 | |
| | $\overline{X} = 70$ | $\overline{X} = 80$ | $\overline{X} = 68$ | |
| | 65 | 92 | 63 | |
| | 81 | 106 | 77 | |
| | 68 | 103 | 70 | |
| | 78 | 95 | 68 | |
| | 70 | 97 | 70 | |
| 90 min | 76 | 101 | 72 | $\overline{X}_{90} = 80.67$ |
| | 72 | 99 | 67 | |
| | 74 | 99 | 73 | |
| | $\overline{X} = 73$ | $\overline{X} = 99$ | $\overline{X} = 70$ | |
| | $\overline{X}_{std} = 71.5$ | $\overline{X}_2 = 89.5$ | $\overline{X}_3 = 69$ | $\overline{\overline{X}} = 76.67$ |

The six combinations of rows and columns are called cells. The six cell means, two row means, three column means, and the grand mean are provided. In two-way ANOVA we want to know if there is a main effect for any of the factors, or, in other words, if there are differences in the means of the levels of one factor averaged across the levels of the other factor. In our example, there are two possible main effects and the essential questions can be phrased as: (1) Is there an overall influence of teaching method (for both the 45-min and 90-min conditions), and (2) is there an overall influence of time devoted to teaching (across all three teaching methods)?

There is an equally, if not more, important question that we can examine with a factorial design: Is there an interaction effect – are the differences among the levels of one factor the same for all levels of the other factor? To better understand this, let us first look at the results of our study in graph form (**Figure 1**).

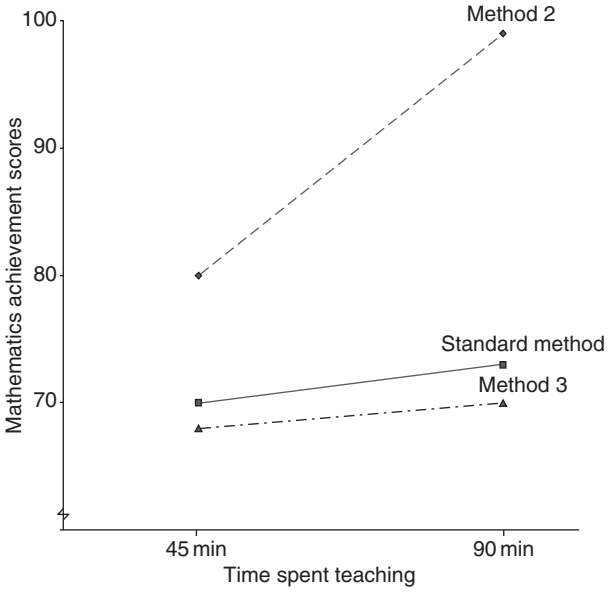


**Figure 1** Hypothetical scores in a 3 × 2 factorial study of the effects of teaching method and time spent teaching on mathematics achievement scores.

In our hypothetical study, if an increase in time spent teaching mathematics had an equal effect for all three teaching methods, the three lines would be parallel. A sizable gap between the lines indicates a possible main effect for teaching method. In our graph, we see evidence for a main effect, but the difference between method 2 and the other two teaching methods depends on the amount of time it is used in the classroom. Method 2 results in moderately better mathematics achievement scores when used for 45 min per day, but dramatically better scores when students are taught mathematics by this method for 90 min per day. If there had been a main effect for teaching method (method 2 better than the other two) without an interaction effect, the differences between method 2 and the others would have been the same for both levels of the other factor (teaching time per day). A significant interaction effect tells us that before we can interpret the effects of a factor, we must examine that factor at each level of the other factor. This will, of course, involve examining cell means. Three-factor experimental designs (e.g., teaching method × time spent teaching × gender) and even higher-order designs are possible, but interpretation of interaction effects becomes more difficult.

When calculating two-way ANOVA, $SS_{total}$ is partitioned into four parts. $SS_{wc}$ (within cell) is the equivalent of $SS_w$ (within groups) in the one-factor design. $SS_{bet}$ is split into three parts: $SS_{rows}$, $SS_{columns}$, and $SS_{rows \times columns}$. To see how to calculate two-way ANOVA, see King and Minium (2008). In our example, we obtain a significant

effect for rows ($F = 30.78$, $df = 1/42$, $p < 0.001$), columns ($F = 80.21$, $df = 2/42$, $p < 0.001$), and rows × columns interaction ($F = 14.59$, $df = 2/42$, $p < 0.001$). As at least one of our main effects was statistically significant, we may now conduct a *post hoc* test to see which levels differed from the others (e.g., method 2 vs. the other two).

## Planned Comparisons

When conducting any statistical test one wants the test to have considerable power, that is, a high probability of rejecting a false null hypothesis. *Post hoc* tests allow us to make all possible pair-wise comparisons, but to protect us from making type I errors, the differences between means must be large enough to be declared statistically significant. If one's study is exploratory in nature, then ANOVA is a very good statistical test. However, if one knows in advance of conducting a study which comparisons are important to him or her, one should use planned (*a priori*) comparisons in place of ANOVA and *post hoc* comparisons (Rosnow and Rosenthal, 1989; Wilkinson *et al.*, 1999; Winer *et al.*, 1991). To do so, one should conduct independent-groups *t* tests using $s^2_w$ (calculated from all groups), but to protect against type I errors, all the comparisons must be orthogonal (unique, with no overlap). When comparing *k* groups, there are only *k*–1 orthogonal comparisons. In our example, we might be interested in the two comparisons standard method versus (method 2 + method 3) and method 2 versus method 3. Unlike *post hoc* comparisons, one does not need a significant overall *F* in order to use planned comparisons. Even Ronald Fisher recognized the value of planned comparisons:

> When the [*F*] test does not demonstrate significant differentiation, much caution should be used before claiming significance for special comparisons. Comparisons, which the experiment was designed to make, may, of course, be made without hesitation (Fisher, 1949: 57).

## Bibliography


Fisher, R. A. (1949). *The Design of Experiments*, 5th edn. London: Oliver and Boyd.

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Research in Education* **42**, 237–288.

King, B. M. and Minium, E. W. (2008). *Statistical Reasoning in the Behavioral Sciences*, 5th edn. Hoboken, NJ: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* **105**, 156–166.

Milligan, G. W., Wong, D. S., and Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin* **101**, 464–470.

Rosnow, R. L. and Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* **44**, 1276–1284.

Tomarken, A. J. and Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin* **99**, 90–99.

Wilkinson, L. and The Task Force on Statistical Inference (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist* **54**, 594–604.

Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical Principles in Experimental Design*, 3rd edn. New York: McGraw Hill.


## Further Reading


Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*, 5th edn. Upper Saddle River, NJ: Prentice Hall.

Gravetter, F. J. and Wallnau, L. B. (2008). *Essentials of Statistics for the Behavioral Sciences*, 6th edn. Belmont, CA: Thomson Wadsworth.

King, B. M. and Minium, E. W. (2008). *Statistical Reasoning in the Behavioral Sciences*, 5th edn. Hoboken, NJ: Wiley.