# Calculating A One-Way Analysis Of Variance

At this point, you should really be patting yourself on the back. You've learned so many different types of descriptive **and** inferential statistics. You've mastered the one-sample *z*-test, the one-sample *t*-test, the independent-samples *t*-test, the dependent-samples *t*-test, and so on. Great work!

While all of the analyses listed above are incredibly useful, however, they all have one limitation in common: They only allow you to compare and contrast between, at most, 2 different groups of people. This is great if you have, say, an experiment with 2 conditions. But in many cases, you might have three or more groups of people that you'd like to compare. If we want to measure whether there are significant differences along some dimension among more than 2 groups of people, we're in trouble if we only have *z*- and *t*-tests to work with. Thankfully, however, we have the one-way analysis of variance (ANOVA). **The ANOVA is a powerful statistical technique because it allows you to measure differences along some dimension between more than 2 groups**. In reality, the ANOVA allows you to measure differences between as many groups as you want! Although it's a bit laborious to calculate by hand, it's a powerful and flexible statistical technique that is well worth our time to learn.

Let's work through an example to really see how the one-way ANOVA shines. I'm a teacher interested in investigating how different study strategies influence grades on a final exam. In particular, I'm interested in three different ways of studying: rote memorization (a learning strategy based primarily on repetition and memorization), verbal practice (in which students talk through course content out loud), and self-testing (which involves self-assessment using, e.g., flash cards). My research question is whether any of these three learning strategies stands out as a very effective (or very ineffective) strategy for studying for a final exam.

To test this research question, I would collect a sample of students that I would randomly assign into 3 different groups that must rely on each of the 3 learning strategies described above: a "*Rote*" group, a "*Verbal*" group, and a "*Test*" group. After having them engage in these learning strategies, I would measure their grade on the final exam (0-100).

Below, I'll show you the data for each of these three groups. The numbers you'll see are scores on the final exam after utilizing each of the 3 learning strategies described above. I'll go ahead and tell you the means for each group: Students who utilized the rote memorization technique earned an average score of $77.8$ on the final exam, those who utilized verbal practice scored an average of $80.9$, and those who actively tested themselves while studying scored an average of $91$. Just using these descriptive statistics, it seems to be the case that testing yourself is a more effective learning strategy than speaking out loud or simply memorizing. But remember, in order to generalize these findings and make conclusions about whether this would hold *in the population* (e.g., among all students), we need to use inferential statistics. Calculating an ANOVA will allow us to do this.

Here's the data:

| Rote | Verbal | Test |
|---|---|---|
| 87 | 89 | 79 |
| 65 | 71 | 92 |
| 76 | 69 | 93 |
| 91 | 92 | 88 |
| 69 | 93 | 89 |
| 78 | 88 | 91 |
| 59 | 64 | 97 |
| 93 | 71 | 88 |
| 88 | 90 | 94 |
| 72 | 82 | 99 |

Okay, so let's figure out how to calculate the one-way ANOVA by hand. I'll note that there are many formulas here, and calculating this involves carefully following a set of steps that, at first, can be laborious. I promise, however, that this becomes relatively simple and formulaic once you get enough practice in! It's not as bad as it looks!

When calculating an ANOVA, our ultimate goal is *F*. ***F* is a test statistic that is simply a tool to allow us to attain a *p*-value; *F* is much like *z* or *t* in this regard**. So, here's the formula to find *F*:

F statistic

$$F = \frac{MS_{between}}{MS_{within}}$$

Notice that we'll need to find $MS_{between}$ and $MS_{within}$ in order to calculate *F*. These terms are called the "Mean Sum of Squares," which are essentially an average Sum of Squares (for between groups and within groups). Unfortunately, each of these terms requires its own formula:

$$MS_{between} = \frac{SS_{between}}{df_{between}} \qquad MS_{within} = \frac{SS_{within}}{df_{within}}$$

And, finally, we need formulas for the Sum of Squares (between groups and within) as well as the degrees of freedom (between groups and within):

$$SS_{between} = \sum \frac{(\sum x)^2}{n} - \frac{(\sum \sum x)^2}{nT} \qquad SS_{within} = \sum \sum (x^2) - \sum \frac{(\sum x)^2}{n}$$

$$df_{between} = k - 1 \qquad df_{within} = nT - k$$

Now, I'll tell you that, despite the overwhelming number of formulas involved in calculating a one-way analysis of variance, most of this problem is actually very quick. It's really just the Sums of Squares ($SS_{between}$, $SS_{within}$) that require a bit of work. Luckily, though, there are a clear set of steps that you can follow to calculate the Sums of Squares in a relatively simple and fool-proof way. (Once you have your Sums of Squares, the problem is 98% complete.) Let's see how to calculate $SS_{between}$ and $SS_{within}$.

Calculating the Sums of Squares involves first finding a few values **within each group** (i.e., separately for each group) and next finding a few values **between groups** (e.g., by adding up across different groups). This process will feel unnatural and contrived at first, but becomes second-nature after practice. Let's go through an example.

## Calculating the Sums of Squares

As a first step in calculating the Sums of Squares (between and within), I would recommend creating a new "squared" column for each group. This simply involves squaring the values in that group and placing the squared values in a new column for each group, much as you did for correlations. I'll show my work for this below:

| Rote | Verbal | Test | Rote_squared | Verbal_squared | Test_squared |
|---|---|---|---|---|---|
| 87 | 89 | 79 | 7569 | 7921 | 6241 |
| 65 | 71 | 92 | 4225 | 5041 | 8464 |
| 76 | 69 | 93 | 5776 | 4761 | 8649 |
| 91 | 92 | 88 | 8281 | 8464 | 7744 |
| 69 | 93 | 89 | 4761 | 8649 | 7921 |
| 78 | 88 | 91 | 6084 | 7744 | 8281 |
| 59 | 64 | 97 | 3481 | 4096 | 9409 |
| 93 | 71 | 88 | 8649 | 5041 | 7744 |
| 88 | 90 | 94 | 7744 | 8100 | 8836 |
| 72 | 82 | 99 | 5184 | 6724 | 9801 |

Notice that the values can get pretty large here. That's okay—it doesn't necessarily mean you made a mistake! Okay, now we're ready to start finding the values we need.

**Steps within each group**

First, we'll need to find the sample size ($n$) of each group. In many cases, $n$ will be the same for each group, but that doesn't have to be the case. Sometimes, you may have more people in one group and fewer in another. This is perfectly fine. In this case, though, each group contains $10$ participants. I'll note that below, with subscripts to keep track of which group we're referring to (e.g., $n_1$ represents the sample size of people in the "*Rote*" group, $n_2$ represents the sample size of people in the "*Verbal*" group, and so on).

$$n_1 = 10 \qquad n_2 = 10 \qquad n_3 = 10$$

Next, we need to find the sum of the values in each group, represented by $\sum x$. You can do this by simply summing each column of data. Doing so will get you the following values:

$$\sum x_1 = 778 \qquad \sum x_2 = 809 \qquad \sum x_3 = 910$$

Now we'll need to find $\frac{(\sum x)^2}{n}$ for each group (i.e., $\frac{(\sum x_1)^2}{n_1}$, $\frac{(\sum x_2)^2}{n_2}$, ,$\frac{(\sum x_3)^2}{n_3}$). This may look ugly but notice that we already calculated $\sum x$ and $n$ for each group! So, finding each of these new terms just involves using the information we already have (i.e., by squaring $\sum x$ and dividing by $n$ for each group). This is how the steps I'm teaching you tend to work; everything builds on what came before it! Okay, let's find these terms:

$$\frac{(\sum x_1)^2}{n_1} = \frac{778^2}{10} = 60528.4 \qquad \frac{(\sum x_2)^2}{n_2} = \frac{809^2}{10} = 65448.1 \qquad \frac{(\sum x_3)^2}{n_3} = \frac{910^2}{10} = 82810$$

The last step for within-group calculations involves taking the sum of the squared values for each group (i.e., finding $\sum(x^2)$ for each group). This is why we started out by squaring each group's values; now, this is a breeze!

$$\sum(x_1^2) = 61754 \qquad \sum(x_2^2) = 66541 \qquad \sum(x_3^2) = 83090$$

Great! Those are all the steps that need to be done **within** each group. Now, let's figure out what we'll need to calculate **between** or **across** groups. You'll quickly see that, after what we did within groups, the between groups calculations will be very easy!

**Steps between groups**

The first between-groups term you'll need to calculate is $nT$ , which is the "total" sample size across all participants in all groups. Think about what this means: Since we already know the sample size for each group ($n_1$, $n_2$, and $n_3$), finding the total sample size across all groups ($nT$) just means adding these values up! Let's do that now:

$$nT = n_1 + n_2 + n_3 = 10 + 10 + 10 = 30$$

The other between-group terms work much like this. The next term we need to find, for example, is $\sum\sum x$. Now, we haven't seen two Sigmas ($\sum$) in a row like this before. What this means is that you're taking the sum of the sum of the values in each group (i.e., the sum of all the Sigma $x$ values, $\sum x$). It's two different sums: first within each group, and then across groups. Since you already found $\sum x$ *within* each group, finding $\sum\sum x$ just means adding those values up! Maybe seeing it will help:

$$\sum\sum x = \sum x_1 + \sum x_2 + \sum x_3 = 778 + 809 + 910 = 2497$$

Next, we'll need to calculate $\frac{(\sum\sum x)^2}{nT}$. Another ugly one, but notice that we already have $nT$ everything we need to find this value! In particular, we know that $nT = 30$ and that $\sum\sum x = 2497$! Finding this new term just means plugging these values in! Make sure to make a note of this value!

$$\frac{(\sum\sum x)^2}{nT} = \frac{2497^2}{30} = 207833.63$$

Next, we'll need to find $\sum \frac{(\sum x)^2}{n}$. Think about what this one means! You already found $\frac{(\sum x)^2}{n}$ for each group (i.e., $\frac{(\sum x_1)^2}{n_1}$, $\frac{(\sum x_2)^2}{n_2}$, $\frac{(\sum x_3)^2}{n_3}$). So, finding $\sum \frac{(\sum x)^2}{n}$ just means adding these three values up! Let's do that now:

$$\sum \frac{(\sum x)^2}{n} = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} = 60528.4 + 65448.1 + 82810 = 208786.5$$

And finally, we'll need to find $\sum\sum(x)^2$. By now you should be seeing a pattern! Do you know what to do to find this term? Since we already have our $\sum(x^2)$ values for each group (i.e., $\sum(x_1^2)$, $\sum(x_2^2)$, and $\sum(x_3^2)$), we only need to add them up to find $\sum\sum(x^2)$. Here we go:

$$\sum\sum(x^2) = \sum(x_1^2) + \sum(x_2^2) + \sum(x_3^2) = 61754 + 66541 + 83090 = 211385$$

### Plugging in

Whew! That's quite a lot of work, right? Thankfully, though, that's 98% of the work! From here on out, we're just plugging in to formulas to find everything we need–no more working with the original data!

Take a look back at the three last terms you calculated: $\frac{(\sum\sum x)^2}{nT}$, $\sum \frac{(\sum x)^2}{n}$, and $\sum\sum(x^2)$. These are 3 terms you need to find your Sums of Squares ($SS_{between}$ and $SS_{within}$). Take a look back at the formulas to see why! We follow these steps to make calculating the Sums of Squares more easily; without the steps, it's even worse!

Let's plug the three values found above into our formulas for $SS_{between}$ and $SS_{within}$:

$$SS_{between} = \sum \frac{(\sum x)^2}{n} - \frac{(\sum\sum x)^2}{nT} = 208786.5 - 207833.63 = 952.87$$

$$SS_{within} = \sum\sum(x^2) - \sum \frac{(\sum x)^2}{n} = 211385 - 208786.5 = 2598.5$$

See how easy that is? Once you follow the steps and find the three terms described above, all you need to do is plug into the formulas for $SS_{between}$ and $SS_{within}$ and then subtract! Now, you're almost done with the problem! All we need to do now is use our Sums of Squares in the remaining formulas displayed on Page 2 of this *Guide*. Following the formulas will lead us all the way to *F*!

## Finding the *F*-test statistic

### Degrees of freedom

First, we'll need to find the degrees of freedom ($df_{between}$ and $df_{within}$), which should only take a moment. This is a simple subtraction with information that you already have (that is, $k$, the number of groups, and $nT$, the total sample size across all groups). Let's take care of the degrees of freedom now:

$$df_{between} = k - 1 = 3 - 1 = 2 \qquad df_{within} = nT - k = 30 - 3 = 27$$

### Working up to *F*

Okay, now we have everything we need to find the *F*-test statistic. It's smooth sailin' from here on out, as we just need to plug the Sums of Squares and degrees of freedom into the formulas presented earlier to work our way back up to *F*.

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{952.87}{2} = 476.43 \qquad MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{2598.5}{27} = 96.24$$

Now that we have our Mean Sums of Squares ($MS_{between}$ and $MS_{within}$), we can finally find *F*.

$$F = \frac{MS_{between}}{MS_{within}} = \frac{476.43}{96.24} = 4.95$$

And that's it! Quite a lot of work, I admit, but it really becomes easy once you get used to it! The bulk of the work involves finding the Sums of Squares. Beyond that, it's just plugging into a variety of formulas!

As a bit of interpretation, I'll tell you that this *F*-test statistic would field a *p*-value of $0.015$, which is smaller than our alpha level of $0.05$. As a result, we would reject the null hypothesis (i.e., we would reject the idea that no differences exist in test performance on the basis of study strategy). It does appear that how you study makes a difference in how well you end up doing on the test.