11.2: Sources of Variance

ANOVA is all about looking at the different sources of variance (i.e. the reasons that scores differ from one another) in a dataset. Fortunately, the way we calculate these sources of variance takes a very familiar form: the Sum of Squares. Before we get into the calculations themselves, we must first lay out some important terminology and notation.

In ANOVA, we are working with two variables, a grouping or explanatory variable and a continuous outcome variable. The grouping variable is our predictor (it predicts or explains the values in the outcome variable) or, in experimental terms, our independent variable, and it made up of k groups, with k being any whole number 2 or greater. That is, ANOVA requires two or more groups to work, and it is usually conducted with three or more. In ANOVA, we refer to groups as "levels", so the number of levels is just the number of groups, which again is k. In the above example, our grouping variable was education, which had 3 levels, so k = 3. When we report any descriptive value (e.g. mean, sample size, standard deviation) for a specific group, we will use a subscript $1\ldots k$ to denote which group it refers to. For example, if we have three groups and want to report the standard deviation s for each group, we would report them as s_1, s_2, and s_3.

Our second variable is our outcome variable. This is the variable on which people differ, and we are trying to explain or account for those differences based on group membership. In the example above, our outcome was the score each person earned on the test. Our outcome variable will still use X for scores as before. When describing the outcome variable using means, we will use subscripts to refer to specific group means. So if we have k = 3 groups, our means will be $\overline{X_1}$, $\overline{X_2}$, and $\overline{X_3}$. We will also have a single mean representing the average of all participants across all groups. This is known as the grand mean, and we use the symbol $\overline{X_G}$. These different means – the individual group means and the overall grand mean – will be how we calculate our sums of squares.

Finally, we now have to differentiate between several different sample sizes. Our data will now have sample sizes for each group, and we will denote these with a lower case "n" and a subscript, just like with our other descriptive statistics: n_1, n_2, and n_3. We also have the overall sample size in our dataset, and we will denote this with a capital N. The total sample size is just the group sample sizes added together.

Between Groups Sum of Squares

One source of variability we can identified in 11.1.3 of the above example was differences or variability between the groups. That is, the groups clearly had different average levels. The variability arising from these differences is known as the between groups variability, and it is quantified using Between Groups Sum of Squares.

Our calculations for sums of squares in ANOVA will take on the same form as it did for regular calculations of variance. Each observation, in this case the group means, is compared to the overall mean, in this case the grand mean, to calculate a deviation score. These deviation scores are squared so that they do not cancel each other out and sum to zero. The squared deviations are then added up, or summed. There is, however, one small difference. Because each group mean represents a group composed of multiple people, before we sum the deviation scores we must multiple them by the number of people within that group. Incorporating this, we find our equation for Between Groups Sum of Squares to be:

$$SS_B = \sum n_j \left(\overline{X}_J - \overline{X_G}\right)^2 \tag{11.2.1}$$

The subscript j refers to the "j^{th}" group where $j = 1\ldots k$ to keep track of which group mean and sample size we are working with. As you can see, the only difference between this equation and the familiar sum of squares for variance is that we are adding in the sample size. Everything else logically fits together in the same way.

Within Groups Sum of Squares

The other source of variability in the figures comes from differences that occur within each group. That is, each individual deviates a little bit from their respective group mean, just like the group means differed from the grand mean. We therefore label this source the Within Groups Sum of Squares. Because we are trying to account for variance based on group-level means, any deviation from the group means indicates an inaccuracy or error. Thus, our within groups variability represents our error in ANOVA.

The formula for this sum of squares is again going to take on the same form and logic. What we are looking for is the distance between each individual person and the mean of the group to which they belong. We calculate this deviation score, square it so that they can be added together, then sum all of them into one overall value:

$$SS_W = \sum \left(X_{ij} - \overline{X}_j \right)^2 \tag{11.2.2}$$

In this instance, because we are calculating this deviation score for each individual person, there is no need to multiple by how many people we have. The subscript j again represents a group and the subscript i refers to a specific person. So, X_{ij} is read as "the i^{th} person of the j^{th} group." It is important to remember that the deviation score for each person is only calculated relative to their group mean: do not calculate these scores relative to the other group means.

Total Sum of Squares

The Between Groups and Within Groups Sums of Squares represent all variability in our dataset. We also refer to the total variability as the Total Sum of Squares, representing the overall variability with a single number. The calculation for this score is exactly the same as it would be if we were calculating the overall variance in the dataset (because that's what we are interested in explaining) without worrying about or even knowing about the groups into which our scores fall:

$$SS_T = \sum \left(X_i - \overline{X_G} \right)^2 \tag{11.2.3}$$

We can see that our Total Sum of Squares is just each individual score minus the grand mean. As with our Within Groups Sum of Squares, we are calculating a deviation score for each individual person, so we do not need to multiply anything by the sample size; that is only done for Between Groups Sum of Squares.

An important feature of the sums of squares in ANOVA is that they all fit together. We could work through the algebra to demonstrate that if we added together the formulas for SS_B and SS_W, we would end up with the formula for SS_T. That is:

$$SS_T = SS_B + SS_W \tag{11.2.4}$$

This will prove to be very convenient, because if we know the values of any two of our sums of squares, it is very quick and easy to find the value of the third. It is also a good way to check calculations: if you calculate each SS by hand, you can make sure that they all fit together as shown above, and if not, you know that you made a math mistake somewhere.

We can see from the above formulas that calculating an ANOVA by hand from raw data can take a very, very long time. For this reason, you will not be required to calculate the SS values by hand, but you should still take the time to understand how they fit together and what each one represents to ensure you understand the analysis itself.
