Partitioning Variability in ANOVA

In ANOVA, the goal is to determine whether there are significant differences among the group means. This is accomplished by partitioning the Total Variation in the response variable (as measured by the corrected total sum of squares) into two components, the Between Group Variation (displayed in the ANOVA table as the Model Sum of Squares) and the Within Group Variation (displayed as the Error Sum of Squares). Comparing the sources of variability enables us to evaluate the null hypothesis. Are the group means equal?

As a high-level example of ANOVA, consider testing the equality of four group means with bell-shaped distributions. You know that the group means are highly different because 1) the center of the distributions is relatively far apart, and 2) the variability within each group is small. This means that, if we sampled data from each group, the sample mean would be highly similar to its true center.

If the within group variability were instead large, and the distributions overlapped, it's possible that the sampled means could be highly similar. Then we are less certain of claiming that the means are significantly different. Thus, in a nut shell, ANOVA simply compares the between group variation to the within group variation to either claim or reject the equality of population means.

Let's look at the three types of sums of squares and how to calculate each using SalePrice as the response and Heating_QC as the predictor. Total Variation, or the Total Sum of Squares, is the overall variability in the response variable. It's calculated as the sum of the squared differences between each observed value and the overall mean.

$$\Sigma\Sigma(Y_{ij} - \overline{\overline{Y}})^2$$

Between Group Variation, or the Model Sum of Squares, is the variability explained by the independent variable and therefore, measures the variability between groups. It's calculated as the weighted sum of the squared differences between the mean for each group and the overall mean.

$$\Sigma n_i(\overline{Y}_i - \overline{\overline{Y}})^2$$

Within Group Variation, or the Error Sum of Squares, is the variability not explained by the model. It's also referred to as within treatment variability or residual sum of squares. Therefore, it measures the random variability within groups. It's calculated as the sum of the squared differences between each observed value and the mean for its group.

$$\Sigma\Sigma(Y_{ij} - \overline{Y}_i)^2$$

The total sum of squares is the sum of the model and error sum of squares. So if you know two of the three types, you can easily calculate the other.