A Quick Introduction to One-Way ANOVA

The following discussion is taken from Moore and McCabe, *Introduction to the Practice of Statistics,* (Freeman, New York, 1993), A one-way analysis of variance (ANOVA) is a significance test of the following sort: From each of several different groups, a simple random sample is taken and the average of each sample is computed. These sample averages are unlikely to be identical. Are their differences mere chance (i.e., the null hypothesis is that all of the averages of the groups are equal), or do they indicate a real difference in the group averages? (If there were only two groups, this would be done by a two-sample z-test; but since there is a distribution of group averages, there are also aspects of a χ^2-test.) As usual, the decision is made by (1) computing from the data a single number, in this case an "F-statistic" (rather than a z-, t- or χ^2-value), (2) consulting the distribution that this statistic will follow if the null hypothesis is true, and (3) accepting the null hypothesis if the computed F-statistic is not too extreme (always "not too large" in this case, because the F-statistic is always positive) and hence not too unlikely, or rejecting it if the F-statistic is too large and unlikely.

In order to say how to compute the F-statistic, we introduce the following notation:

I = number of groups (running variable g)
n_g = number in sample from group g
$N = n_1 + n_2 + \cdots + n_I$, total number of data values
$x_{i,g}$ = i-th data value in sample from group g
(g runs from 1 to I, i runs from 1 to n_g)
$\overline{x_g} = (\sum_{i=1}^{n_g} x_{i,g})/n_g$, the average of the sample from group g
$\overline{x} = (\sum_{g=1}^{I} \sum_{i=1}^{n_g} x_{i,g})/N$, the average of all the data values.

Note that

$$\begin{aligned}
&\sum_{g=1}^{I}\sum_{i=1}^{n_g}(x_{i,g}-\overline{x_g})^2 + \sum_{g=1}^{I}\sum_{i=1}^{n_g}(\overline{x_g}-\overline{x})^2) - \sum_{g=1}^{I}\sum_{i=1}^{n_g}(x_{i,g}-\overline{x})^2 \\
&\quad = \sum_{g=1}^{I}\sum_{i=1}^{n_g}(x_{i,g}^2 - 2x_{i,g}\overline{x_g} + \overline{x_g}^2 + \overline{x_g}^2 - 2\overline{x_g x} + \overline{x}^2 - x_{i,g}^2 + 2x_{i,g}\overline{x} - \overline{x}^2) \\
&\quad = 2\sum_{g=1}^{I}\sum_{i=1}^{n_g}(x_{i,g}\overline{x} - x_{i,g}\overline{x_g} + \overline{x_g}^2 - \overline{x_g x}) = 2\sum_{g=1}^{I}\sum_{i=1}^{n_g}(\overline{x}-\overline{x_g})(x_{i,g}-\overline{x_g}) \\
&\quad = 2\sum_{g=1}^{I}(\overline{x}-\overline{x_g})\left(\left(\sum_{i=1}^{n_g} x_{i,g}\right) - n_g\overline{x_g}\right) = 2\sum_{g=1}^{I}(\overline{x}-\overline{x_g})(0) = 0
\end{aligned}$$

(Moore and McCabe do not show this computation; they say only, "It is an algebraic fact that ...". Now we know why.) We set:

$$SSE = \sum_{g=1}^{I}\sum_{i=1}^{n_g}(x_{i,g}-\overline{x_g})^2 \qquad \text{and} \qquad SSG = \sum_{g=1}^{I}\sum_{i=1}^{n_g}(\overline{x_g}-\overline{x})^2 = \sum_{g=1}^{I} n_g(\overline{x_g}-\overline{x})^2 \ .$$

The differences $x_{i,g} - \overline{x_g}$ are the deviations of the data values from the average of their group's sample; SSE abbreviates the "sum of square deviations due to error" (though "residual" is often a better term than "error"). The differences $\overline{x_g} - \overline{x}$ are the deviations of the group samples' averages from the overall average; SSG means "sum of square deviations due to groups". The point of the

computation above is that $\sqrt{(SSE + SSG)/N}$ is the SD of the pooled sample, i.e., all the group samples thrown together and treated as a single sample of the total population.

If the null hypothesis is true, then we should expect SSE (i.e., the variations within the groups) to be large relative to SSG (the variations of the sample averages from the grand average) — the null hypothesis is that the averages of the groups are all the same, so the expected values of the $x_{i,g}$'s and the $\overline{x_g}$'s are all this common average, but there should be less variability in the averages $\overline{x_g}$'s. Rather than taking their ratio immediately, however, we first divide SSE and SSG by their respective degrees of freedom, $DFE = N - I$ and $DFG = I - 1$, to get the MSE ("mean square deviation due to error") and MSG ("mean square deviation due to groups"). Then

$$MSE = \frac{SSE}{DFE} \ , \qquad MSG = \frac{SSG}{DFG} \ , \qquad F = \frac{MSG}{MSE} \ .$$

If F is not too large, we accept the null hypothesis; if F is large enough, we reject it. And, as usual, "large enough" is determined by consulting an F-table of probabilities, indexed this time by both degrees of freedom, DFE and DFG. (To keep the table within the two dimensions of a page, therefore, only the F-value large enough to give a 5% significance level appears in the table; but if other pages of tables are available, they may give this information for significance cutoff levels other than 5%, i.e., for other "alpha levels".)

Finally, we note some assumptions that must hold for the F-distribution to be correct in ths situation: Namely, the groups from which the simple random samples are taken must each follow the normal curve, with (1) possibly different population averages (though the null hypothesis is that all are equal), and (2) all the same population standard deviation. Condition (2) cannot be checked from the data, but Moore and McCabe assure us that it is sufficient to have the largest of SD's of the group samples not be more than twice the smallest.