1 Analysis of Variance versus Experimental Design

Not all data analyzed by ANOVA are from a designed experiment. On the other hand, some designed experiments lead to data for which ANOVA methods are inappropriate. However, there is a strong historical connection between ANOVA and Experimental Design.

2 Analysis of Variance

The key idea behind an analysis of variance involves a decomposition of the total sum of squares. For a one-way ANOVA (possibly arising from a completely randomized design) the decomposition is TSS = SSB + SSW (Our text uses SST instead of SSB, and SSE instead of SSW). This expression not only looks a lot like TSS = SSR + SSE from regression, but we will see later that we can set up a regression model using dummy variables for which these expressions are identical, where SSB = SSR and SSW = SSE.

If y_{ij} is the jth observation in group i, then $y_{ij} - \overline{y}_{..} = (y_{ij} - \overline{y}_{i.}) + (\overline{y}_{i.} - \overline{y}_{..})$, which leads to:

$$\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n}(\overline{y}_{i.} - \overline{y}_{..})^2, \text{ or TSS = SSW +SSB.}$$

As an example, consider three groups with the following data: Group 1 has y_{1j} values of 1, 2, and 3, Group 2 has y_{2j} values of 5, 3, and 4, and Group 3 has y_{3j} values of 6, 7, and 5. The overall sample mean is $\overline{y}_{..} = (\sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij})/tn = 36/9 = 4$. Then TSS is

$$\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2 = (1-4)^2 + (2-4)^2 + ... + (5-4)^2 = 30.$$

The group means are $\overline{y}_{1.} = 2, \overline{y}_{2.} = 4$, and $\overline{y}_{3.} = 6$, so SSW and SSB are

$$\text{SSW} = \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 = (1-2)^2 + (2-2)^2 + (3-2)^2 + (5-4)^2 + ... + (5-6)^2 = 6, \text{ and}$$

$$\text{SSB} = \sum_{i=1}^{t}\sum_{j=1}^{n}(\overline{y}_{i.} - \overline{y}_{..})^2 = \sum_{i=1}^{t} n(\overline{y}_{i.} - \overline{y}_{..})^2 = 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24.$$

Thus TSS = SSW +SSB or 30 = 6 + 24 partitions the total sum of squares about the overall mean into two parts, one within groups (due to error, or effects not accounted by the model) and one between groups (measuring the difference between sample means). Since each group here has n observations, each group contributes $n-1$ degrees of freedom for the within group sum of squares, for a total of $t(n-1)$ degrees of freedom for SSW. SSB is calculating the sum of squares of t sample means about their (overall) mean, so it has $t-1$ degrees of freedom. For the example data above, $t(n-1) = 3(2) = 6$, and $t-1 = 3$ - 1 = 2. We can summarize this information in an analysis of variance table:

Source	SS	df	MS	F
Between groups	24	2	12	12
Within groups	6	6	1	
Total sum of squares	30	8		

To test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative hypothesis H_a : some μ_i's differ, we compare the F statistic to an F distribution with numerator df = $t - 1 = 3$ - 1 = 2, and denominator df = $t(n - 1) = 3(2) = 6$. When group sample sizes are unequal we replace n in the expressions by n_i, which is the sample size in the ith group.

3 The model for a completely randomized experiment (1 way ANOVA)

The model for ANOVA with one grouping factor is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where μ is the population grand mean, $\alpha_i = \mu_i - \mu$ is the treatment effect for the ith group, and $\varepsilon_{ij} = y_{ij} - \mu - \alpha_i$ is the random error effect for y_{ij}. Using this notation we can write the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_t$ in the alternate form $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_t = 0$. In performing ANOVA for a completely randomized experiment we assume that 1) random samples have been taken from each of the t populations, 2) the errors ε_{ij} have a normal distribution with mean 0, and 3) the errors ε_{ij} have a common variance σ^2. It can be shown that $E(MSW) = \sigma^2$, and $E(MSB) = \sigma^2 + n \sum \alpha_i^2/(t - 1)$. Then if H_0 is true, all α_i terms equal zero and MSW and MSB be nearly equal. Thus $F = MSB/MSW \approx 1$ when H_0 is true and $F > 1$ when H_0 is false.