1 Completely Randomized Design (CRD) Part I

It is the simplest of designs, but even planning for it requires scientific and statistical decisions (Acid rain example).

1.1 Exploratory Data Analysis (EDA)

Essential for any statistical analysis. For the CRD we often use boxplots for initial examination of the data. From the boxplot we can look for evidence of a treatment difference, and possible outliers or problems with homogeneity of variance.

1.2 ANOVA as a choice of the best fitting model for the mean

Let y_{ij} $(i = 1, ..., g;\ j = 1, ..., n_i)$ be the jth observation in group i. In ANOVA from a CRD we consider two models for y_{ij}. The first model, $y_{ij} = \mu_i + \varepsilon_{ij}$, specifies that each group has a different mean value μ_i. This is also called the **full** model for y_{ij}. The second model, $y_{ij} = \mu + \varepsilon_{ij}$, specifies that all groups have identical mean values μ. This is also called the **reduced** model for y_{ij}. Note that the reduced model is a special case (or subset) of the full model. Both models make the assumption that the ε_{ij} are independent, have mean zero, and variance σ^2. To conduct statistical inference (tests, confidence intervals, etc.) we make the further assumption that the ε_{ij} have a normal distribution.

An alternative way to express the models above is by letting $\mu_i = \mu^* + \alpha_i$, where μ^* is the overall mean and α_i is the treatment effect of group i. Then the full model is $y_{ij} = \mu^* + \alpha_i + \varepsilon_{ij}$. This new formulation of the full model generalizes well to more complicated models, but introduces a complication because there are now more parameters than groups. For both the full and reduced models, we can develop estimators for the parameters μ and σ^2 (reduced) or μ, μ_i, α_i, and σ^2 (full), shown in Display 3.1 in the text. We can also calculate confidence intervals for our parameters.

To choose between the full and reduced models, we compare their sum of squared residuals (SSR). A residual r is the error in predicting an observation, $r = y - \widehat{y}$, where $\widehat{y}$ is the predicted value of y. For the full model, $\widehat{y}_{ij} = \overline{y}_{i\cdot}$ (the sample group mean), and for the reduced model, $\widehat{y}_{ij} = \overline{y}_{\cdot\cdot}$ (the sample overall mean). SSR for the full model can never exceed SSR for the reduced model, so we wish to decide if SSR for the full model has been reduced enough to account for the extra parameters (μ_i) in the full model.

1.3 Analysis of Variance mechanics

Analysis of variance involves a partition of the total sum of squares for the observations y_{ij}. Using our notation from above, $y_{ij} - \overline{y}_{\cdot\cdot} = (y_{ij} - \overline{y}_{i\cdot}) + (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})$, which equals the residual (from the full model) plus the ith treatment effect, or $r_{ij} + \widehat{\alpha}_i$. By squaring and summing these terms and cancelling the cross product we obtain:

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\cdot\cdot})^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2, \text{ or } \mathrm{SS}_T = \mathrm{SS}_E + \mathrm{SS}_{Trt}.$$

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 $\overline{y}_{\cdot\cdot} = (\sum_{i=1}^{g}\sum_{j=1}^{3} y_{ij})/3g = 36/9 = 4$. Then SS_T is

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

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

$$\mathrm{SS}_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2 = (1-2)^2+(2-2)^2+(3-2)^2+(5-4)^2+...+(5-6)^2 = 6, \text{ and}$$

$$\mathrm{SS}_{Trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i.}-\overline{y}_{..})^2 = \sum_{i=1}^{g} n_i(\overline{y}_{i.}-\overline{y}_{..})^2 = 3(2-4)^2+3(4-4)^2+3(6-4)^2 = 24.$$

Thus $\mathrm{SS}_T = \mathrm{SS}_E + \mathrm{SS}_{Trt}$ 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 (due to the treatment, measuring the difference between sample means). Since each group here has n_i observations, each group contributes $n_i - 1$ degrees of freedom for the error sum of squares, for a total of $\sum(n_i - 1)$ degrees of freedom for SS_E. SS_{Trt} is calculating the sum of squares of g sample means about their (overall) mean, so it has $g - 1$ degrees of freedom. For the example data above, $\sum(n_i - 1) = g(n - 1)$ $= 3(2) = 6$, and $g - 1 = 3 - 1 = 2$. 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		

Three distinct ways to consider these calculations are: 1) Focus attention only on the partition of SS_T, $30 = 6 + 24$, 2) To test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative hypothesis H_a : some μ_i's differ, using an F test, or 3) View it as a model selection problem, comparing SS_T (= SSR for reduced model) to SS_E (= SSR for the full model) to decide the best fitting model.