7

Multifactor Designs

The analysis of randomized blocks can be generalized to very complex experimental designs with multiple control variables and confounded effects. In this chapter, we consider the evaluation of main effects and interactions via synchronized rearrangements and the analysis of variance. We also study the analysis of covariance and the analysis of unbalanced designs via a combination of bootstrap and permutation methods.

7.1 Multifactor Models

What distinguishes the complex experimental design from the simple one-sample, two-sample, and k-sample experiments we have considered so far is the presence of multiple control factors. For example, we may want to assess the simultaneous effects on crop yield of hours of sunlight and rainfall. We determine to observe the crop yield X_{ijm} for I different levels of sunlight, $i = 1, \ldots, I$, and J different levels of rainfall, $j = 1, \ldots, J$, and to make M observations at each factor combination, $m = 1, \ldots, M$. We adopt as our model relating the dependent variable crop yield (the *effect*) to the independent variables of sunlight and rainfall (the *causes*)

$$X_{ijm} = \mu + s_i + r_j + (sr)_{ij} + \epsilon_{ijm}. \tag{7.1}$$

In this model, terms with a single subscript like s_i, the effect of sunlight, are called *main effects*. Terms with multiple subscripts like $(sr)_{ij}$, the residual and nonadditive effect of sunlight and rainfall, are called *interactions*. The $\{\epsilon_{ijm}\}$ represent that portion of crop yield that cannot be explained by the independent variables alone; these are variously termed the *residuals*, the *errors*, or the *model errors*. To ensure the residuals are exchangeable so that permutation methods can be applied, the experimental units must be assigned at random to treatment.

If we wanted to assess the simultaneous effect on crop yield of three factors simultaneously—sunlight, rainfall, and fertilizer, say—we would observe the crop yield X_{ijkm} for I different levels of sunlight, J different levels of rainfall, and K different levels of fertilizer, $k = 1, \ldots, K$, and make n_{ijk} observations at each factor combination. Our model would then be

$$X_{ijkm} = \mu + s_i + r_j + f_k + (sr)_{ij} + (sf)_{ik} + (rf)_{jk} + (srf)_{ijk} + \epsilon_{ijkm}. \quad (7.2)$$

In this model we have three main effects, s_i, r_j and f_k, three first-order interactions, $(sr)_{ij}, (sf)_{ik}$, and $(rf)_{jk}$, a single second-order interaction, $(srf)_{ijk}$, and the error term, ϵ_{ijkm}.

Including the additive constant μ in the model allows us to define all main effects and interactions so they sum to zero; thus $\sum s_i = 0, \sum r_j = 0$, $\sum_i (sr)_{ij} = 0$ for $j = 1, \ldots, J$, $\sum_j (sr)_{ij} = 0$ for $i = 1, \ldots, I$, and so forth. Under the hypothesis of no nonzero interactions, the expected effect of the joint presence of the two factors s_i and r_j is the sum $s_i + r_j$. Under the hypothesis of "no effect of sunlight on crop yield," each of the main effects are equal, that is $s_1 = \cdots = s_I = 0$. Under one alternative to this hypothesis, the different terms s_i represent deviations from a zero average.

Clearly, when we have multiple factors, we must also have multiple test statistics. In the preceding example, we require three separate tests and test statistics for the three main effects of sunlight, rainfall, and fertilizer, plus four other statistical tests for the three first-order and the one second-order interactions. Will we be able to find statistics that measure a single intended effect without confounding it with a second unrelated effect? Will the several p-values be independent of one another?

7.2 Analysis of Variance

The analysis of variance (ANOVA) relies on the decomposition (that is, the analysis) of the sum of squares of a set of observations about their grand mean into a series of sums. For example in the two-way ANOVA,

$$\begin{aligned}\sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{...})^2 &= \sum_i n_{i.}(\bar{X}_{i..} - \bar{X}_{...})^2 + \sum_j n_{.j}(\bar{X}_{.j.} - \bar{X}_{...})^2 \\ &\quad + \sum_i \sum_j n_{ij}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2 \\ &\quad + \sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{ij.})^2.\end{aligned}$$

Associated with the model of Equation (7.1) are a set of univariate linear hypotheses concerning the main effects $\{s_i\}$ and $\{r_j\}$ and the interactions $(sr)_{ij}$.