Two Factor Studies
Ed Stanek

Introduction

We develop the background for analysis of variance tables for balanced two factor studies. We first consider studies that are completely randomized.

Study Population and Parameters

We assume the study is conducted using a large population of \( s = 1, \ldots, N \) subjects. Potentially, each subject could receive any combination of levels of two factors. We refer to the factors as A and B, where there are \( i = 1, \ldots, a \) levels of factor A and \( j = 1, \ldots, b \) levels of factor B. There are a total of \( ab \) factor combinations. Each subject could potentially receive any of the factor combinations. We represent the potentially observable response of a subject after treatment as a fixed constant \( \mu_{ijs} \). This representation assumes that there is no response error.

Using this representation of response, we define parameters for the mean of various treatment levels as

\[
\begin{align*}
\mu_i &= \frac{1}{bN} \sum_{j=1}^{b} \sum_{s=1}^{N} \mu_{ijs} ,
\mu_j &= \frac{1}{aN} \sum_{i=1}^{a} \sum_{s=1}^{N} \mu_{ijs} ,
\mu_{ij} &= \frac{1}{N} \sum_{s=1}^{N} \mu_{ijs} \quad \text{and} \quad \mu = \frac{1}{abN} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{s=1}^{N} \mu_{ijs} .
\end{align*}
\]

With these definitions, we define the parameters \( \alpha_i = \mu_i - \mu \), for \( i = 1, \ldots, a \) and \( \beta_j = \mu_j - \mu \) for \( j = 1, \ldots, b \), and \( \gamma_{ij} = \left( \mu_{ijs} - \mu_{i..} \right) - \left( \mu_{..j} - \mu \right) \) so that

\[
\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}
\]

and

\[
\mu_{ijs} = \mu_{ij} + e_{ijs}
\]

where \( e_{ijs} = \mu_{ijs} - \mu_{ij} \). We define expressions for the variance as

\[
\left( \frac{N-1}{N} \right)^2 \sigma_i^2 = \left( \frac{N-1}{N} \right) \left( \sum_{s=1}^{N} \left( \mu_{ijs} - \mu_{ij} \right)^2 \right) \frac{N-1}{N-1}
\]

for \( i = 1, \ldots, a \) and \( j = 1, \ldots, b \).

Simplifying Assumptions

We simplify the problem with some additional assumptions. Let us assume that:

1. \( N \) is very large so that we can approximate \( \frac{N-1}{N} \approx 1 \).
2. The variance is the same for all factor level combinations, \( \sigma^2_{ij} = \sigma^2 \) for \( i = 1, \ldots, a \) and \( j = 1, \ldots, b \).

3. There is a constant subject effect (for any factor combination), such that \( e_{ij} = \mu_{ij} - \mu_i = \delta_i \) for all \( i = 1, \ldots, a \) and \( j = 1, \ldots, b \).

**The Study**

The study is conducted by randomly permuting the population of \( N \) subjects, and then partitioning the permutation into consecutive groups of \( n \) units. For each group of units, we assign a factor level combination, and then observe response. Factor level combinations are assigned such that the first group of units in the permutation receives \( i = 1, j = 1 \); the second group receives \( i = 1, j = 2 \); etc.

**Simplifying Assumptions**

We make the additional assumption that:

4. There are no replications, such that \( n = 1 \)

5. The population size is large relative to the total study size, such that \( N \gg ab \).

**The Model for Random Variables**

We refer to the potential response for units in a randomly selected permutation as random variables. The response is random since only one permutation is selected. We refer to the units in a permutation by their position. Since factor levels are assigned to positions, we use the factor levels to specify the positions. We index the units (within a factor level combination) with the sub-script \( k \), and note that as a result of assumption 4, only one unit receives each factor level combination and hence \( k = 1 \). Since \( k \) is always one, we drop the subscript from our notation.

We represent the potentially observable response for unit \( k \) that receives the \( ij \) factor level combination as

\[
Y_{ijk} = \mu_{ij} + E_{ijk}
\]

where \( E(Y_{ijk}) = \mu_i \), and (using assumptions 2 and 3), \( \text{var}(Y_{ijk}) = \sigma^2 \). Using expressions for a random permutation of the population, we note that the study response vector (of dimension \( ab \times 1 \)) can be represented as

\[
Y = \mu + E
\]

where \( Y = (Y_{11}, Y_{12}, \ldots, Y_{ab})' \), \( \mu = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ib})' \) and \( E = (E_{111}, E_{121}, \ldots, E_{ab1})' \),

where \( E(Y) = \mu \), and \( \text{var}(Y) = \sigma^2 \left( I_{ab} - \frac{J_{ab}}{N} \right) \). Note that we can represent
\[ \mu = \left( X_1 \mid X_2 \mid X_3 \mid X_4 \right) \begin{pmatrix} \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix} \] where \( X_1 = I_{a b} \), \( X_2 = \left( \frac{I_{a-1}}{-I_{a-1}} \right) \otimes I_{b} \), \( X_3 = I_{a} \otimes \left( \frac{I_{b-1}}{-I_{b-1}} \right) \), and

\[ X_4 = \left( \frac{I_{a-1}}{-I_{a-1}} \right) \otimes \left( \frac{I_{b-1}}{-I_{b-1}} \right) \].

We represent this equivalently as \( \mu = X\beta \) where

\[ X = \left( X_1 \mid X_2 \mid X_3 \mid X_4 \right) \), \( \alpha = \left( \alpha_1 \alpha_2 \cdots \alpha_{a-1} \right) \), \( \beta = \left( \beta_1 \beta_2 \cdots \beta_{b-1} \right) \) and

\[ \gamma = \left( \gamma_1 \gamma_2 \cdots \gamma_{a-1} \right) \) where \( \gamma_i = \left( \gamma_{i1} \gamma_{i2} \cdots \gamma_{i(b-1)} \right) \).

Using this notation, and making use of assumption 5, the model is given by

\[ Y = X\beta + E \]

where \( E(Y) = X\beta \), and \( \text{var}(Y) = V = \sigma^2I_{ab} \).

**Estimation and the ANOVA Table**

Since \( E(Y) = X\beta \) and \( \text{var}(Y) = V = \sigma^2I_{ab} \), estimates of \( \beta \) can be developed by specifying a criteria, and then deriving an estimator that optimizes the criteria. We require the estimator to be linear in the sample data, unbiased (such that \( E(\hat{\beta}) = \beta \)) and to minimize the quantity

\[ E'V^{-1}E = (Y - X\beta)'V^{-1}(Y - X\beta). \]

The resulting estimator is \( \hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \), which reduces to \( \hat{\beta} = (X'X)^{-1}X'Y \) when given by \( V = \sigma^2I_{ab} \).

Using the estimator, we can partition the total sums of squares given by

\[ Y'Y = (Y - X\hat{\beta})'Y + \hat{\beta}'X'X\hat{\beta}. \] The first term in this expression corresponds to the residual sums of squares. Let us represent \( H_x = X(X'X)^{-1}X' \) for any given matrix \( X \). Then

\[ (Y - X\hat{\beta})'Y + \hat{\beta}'X'X\hat{\beta} = Y'(I - H_x)Y. \] The second term in the expression simplifies to

\[ \hat{\beta}'X'X\hat{\beta} = Y'X(X'X)^{-1}X'Y. \] Then noting that all the sub-matrices in \( X = \left( X_1 \mid X_2 \mid X_3 \mid X_4 \right) \) are orthogonal, \( H_x = H_{x_1} + H_{x_2} + H_{x_3} + H_{x_4} \), and

\[ \hat{\beta}'X'X\hat{\beta} = Y'H_{x_1}Y + Y'H_{x_2}Y + Y'H_{x_3}Y + Y'H_{x_4}Y. \]

Combining these terms, we find that

\[ Y'Y = Y'(I - H_x)Y + Y'H_{x_1}Y + Y'H_{x_2}Y + Y'H_{x_3}Y + Y'H_{x_4}Y \]

or equivalently, that
These are the terms that are given in a two-way ANOVA table.

Table 1. ANOVA Table for Two-Factor Problem with Corrected SS

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Sq.</th>
<th>DF</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor A</td>
<td>Y'(H_{x_1})Y</td>
<td>a-1</td>
<td>Y'(H_{x_1})Y / (a-1)</td>
<td></td>
</tr>
<tr>
<td>Factor B</td>
<td>Y'(H_{x_2})Y</td>
<td>b-1</td>
<td>Y'(H_{x_2})Y / (b-1)</td>
<td></td>
</tr>
<tr>
<td>A * B</td>
<td>Y'(H_{x_1})Y</td>
<td>(a-1)(b-1)</td>
<td>Y'(H_{x_1})Y / (a-1)(b-1)</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>Y'(I - H_{x_1})Y</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>Y'(I - H_{x_1})Y</td>
<td>ab-1</td>
<td>Y'(I - H_{x_1})Y / ab-1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Y'Y</td>
<td>ab</td>
<td>Y'Y / ab</td>
<td></td>
</tr>
</tbody>
</table>

Using Cochran's Theorem, we can evaluate the EMS.