4  FACTORIAL DESIGNS

4.1  Two Factor Factorial Designs

- A two-factor factorial design is an experimental design in which data is collected for all possible combinations of the levels of the two factors of interest.

- If equal sample sizes are taken for each of the possible factor combinations then the design is a balanced two-factor factorial design.

- A balanced \(a \times b\) factorial design is a factorial design for which there are \(a\) levels of factor \(A\), \(b\) levels of factor \(B\), and \(n\) independent replications taken at each of the \(a \times b\) treatment combinations. The design size is \(N = abn\).

- The effect of a factor is defined to be the average change in the response associated with a change in the level of the factor. This is usually called a main effect.

- If the average change in response across the levels of one factor are not the same at all levels of the other factor, then we say there is an interaction between the factors.

\[
\begin{align*}
\text{TYPE} & \quad \\ \text{Cell}(i,j) & \quad \text{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk} \\
\text{i}^{th} \text{ level of } A & \quad y_{i.} = \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\text{j}^{th} \text{ level of } B & \quad y_{.j} = \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\text{Overall} & \quad y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}
\end{align*}
\]

where \(n_{ij}\) is the number of observations in cell \((i,j)\).

**EXAMPLE (A \(2 \times 2\) balanced design):** A virologist is interested in studying the effects of \(a = 2\) different culture media \((M)\) and \(b = 2\) different times \((T)\) on the growth of a particular virus. She performs a balanced design with \(n = 6\) replicates for each of the \(4 M \times T\) treatment combinations. The \(N = 24\) measurements were taken in a completely randomized order. The results:

<table>
<thead>
<tr>
<th>THE DATA</th>
<th>TOTALS</th>
<th>MEANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M)</td>
<td>(T = 1)</td>
<td>(T = 2)</td>
</tr>
<tr>
<td>12 T hours</td>
<td>(Y_{11.} = 140)</td>
<td>(Y_{12.} = 156)</td>
</tr>
<tr>
<td>18 T hours</td>
<td>(Y_{21.} = 223)</td>
<td>(Y_{22.} = 192)</td>
</tr>
<tr>
<td>Medium 1</td>
<td>(Y_{1.} = 363)</td>
<td>(Y_{2.} = 348)</td>
</tr>
<tr>
<td>Medium 2</td>
<td>(Y_{..} = 711)</td>
<td></td>
</tr>
</tbody>
</table>

- For medium 1, the \(T\) effect = \(37.16 - 23.3\) =
- For medium 2, the \(T\) effect = \(32 - 26\) =

- The effect on the response of changing \(M\) from medium 1 to 2 depends on the level of \(T\).

- For \(T = 12\) hours, the \(M\) effect = \(26 - 23.3\) =
- For \(T = 18\) hours, the \(M\) effect = \(32 - 37.16\) =

125
• If either of these pairs of estimated effects are significantly different then we say there exists a **significant interaction** between factors $M$ and $T$. For the $2 \times 2$ design example:
  - If $13.83$ is significantly different than $6$ for the $M$ effects, then we have a significant $M \times T$ interaction.
  Or,
  - If $2.6$ is significantly different than $-5.16$ for the $T$ effects, then we have a significant $M \times T$ interaction.

• There are two ways of defining an interaction between two factors $A$ and $B$:
  - If the average change in response between the levels of factor $A$ is not the same at all levels of factor $B$, then an interaction exists between factors $A$ and $B$.
  - The lack of additivity of factors $A$ and $B$, or the nonparallelism of the mean profiles of $A$ and $B$, is called the interaction of $A$ and $B$.

• When we assume there is no interaction between $A$ and $B$, we say the effects are **additive**.

• An **interaction plot** or **treatment means plot** is a graphical tool for checking for potential interactions between two factors. To make an interaction plot,
  1. Calculate the cell means for all $a \cdot b$ combinations of the levels of $A$ and $B$.
  2. Plot the cell means against the levels of factor $A$.
  3. Connect and label means the same levels of factor $B$.

• The roles of $A$ and $B$ can be reversed to make a second interaction plot.

• Interpretation of the interaction plot:
  - Parallel lines usually indicate no significant interaction.
  - Severe lack of parallelism usually indicates a significant interaction.
  - Moderate lack of parallelism suggests a possible significant interaction may exist.

• Statistical significance of an interaction effect depends on the magnitude of the $MS_E$:
  For small values of the $MS_E$, even small interaction effects (less nonparallelism) may be significant.

• When an $A \times B$ interaction is large, the corresponding main effects $A$ and $B$ may have little practical meaning. Knowledge of the $A \times B$ interaction is often more useful than knowledge of the main effect.

• We usually say that a significant interaction can mask the interpretation of significant main effects. That is, the experimenter must examine the levels of one factor, say $A$, at fixed levels of the other factor to draw conclusions about the main effect of $A$.

• It is possible to have a significant interaction between two factors, while the main effects for both factors are not significant. This would happen when the interaction plot shows interactions in different directions that balance out over one or both factors (such as an $X$ pattern). This type of interaction, however, is uncommon.
4.2 The Interaction Model

- The interaction model for a two-factor completely randomized design is:

\[ y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad (22) \]

where \( \mu \) is the baseline mean, \( \alpha_i \) is the \( i^{th} \) factor \( A \) effect, \( \beta_j \) is the \( j^{th} \) factor \( B \) effect, \( (\alpha \beta)_{ij} \) is the \( (i,j)^{th} \) \( A \times B \) interaction effect, and \( \epsilon_{ijk} \) is the random error of the \( k^{th} \) observation from the \( (i,j)^{th} \) cell.

We assume \( \epsilon_{ijk} \sim IID N(0, \sigma^2) \). For now, we will also assume all effects are fixed.

- If \( (\alpha \beta)_{ij} \) is removed from (22), we would have the additive model:

\[ y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad (23) \]

- If we impose the constraints

\[ \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0 \quad \sum_{i=1}^{a} (\alpha \beta)_{ij} = 0 \quad \text{for all } j \quad \text{and} \quad \sum_{j=1}^{b} (\alpha \beta)_{ij} = 0 \quad \text{for all } i \quad (24) \]

then the least squares estimates of the model parameters are

\[ \hat{\mu} = \hat{\alpha}_i = \hat{\beta}_j = \hat{\alpha \beta}_{ij} = \]

- If we substitute these estimates into (22) we get

\[ y_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha \beta}_{ij} + \epsilon_{ijk} \]

\[ = \bar{y}_{..} + (\bar{y}_{i..} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{ij} - \bar{y}_{i..} - \bar{y}_{.j} + \bar{y}_{..}) + \epsilon_{ijk} \]

where \( \epsilon_{ijk} \) is the \( k^{th} \) residual from the treatment \( (i,j)^{th} \) cell, and \( \epsilon_{ijk} = \)

- For the \( 2 \times 2 \) design,

\[ \bar{y}_{..} = 29.625 \quad \bar{y}_{1.} = 24.6 \quad \bar{y}_{2.} = 34.586 \quad \bar{y}_{.1} = 30.25 \quad \bar{y}_{.2} = 29.00 \]

- Assuming the constraints in (24),

\[
\begin{align*}
\alpha_1 &= 24.6 - 29.625 = \\
\alpha_2 &= 34.583 - 29.625 = \\
\beta_1 &= 30.256 - 29.625 = \\
\beta_2 &= 29.006 - 29.625 = \\
\alpha \beta_{11} &= 23.3 - 24.6 - 30.25 + 29.625 = \\
\alpha \beta_{12} &= 26 - 24.6 - 29.00 + 29.625 = \\
\alpha \beta_{21} &= 37.16 - 34.583 - 30.25 + 29.625 = \\
\alpha \beta_{22} &= 32 - 34.583 - 29.00 + 29.625 = \]
\]

127
Estimation of Parameters

Model: \( y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij} \)

Goal: Minimize \( L = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - (\hat{\alpha}\hat{\beta})_{ij})^2 \)

\[
\begin{align*}
(1) & \quad \hat{\mu} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})}{ab} \\
(2) & \quad \hat{\alpha}_i = \frac{\sum_{j=1}^{b} (y_{ij} - \bar{y}_{i..})}{b} \\
(3) & \quad \hat{\beta}_j = \frac{\sum_{i=1}^{a} (y_{ij} - \bar{y}_{..j})}{a} \\
(4) & \quad \hat{\alpha}\hat{\beta}_{ij} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})}{(a-1)(b-1)} \\
\end{align*}
\]

Constraints: \( \sum_{i=1}^{a} \hat{\alpha}_i = 0 \), \( \sum_{j=1}^{b} \hat{\beta}_j = 0 \), \( \sum_{i=1}^{a} \sum_{j=1}^{b} (\hat{\alpha}\hat{\beta})_{ij} = 0 \), \( \sum_{j=1}^{b} (\hat{\alpha}\hat{\beta})_{ij} = 0 \)

From (1) \( a b n \hat{\mu} = \bar{y}_{..} \rightarrow \boxed{\hat{\mu} = \bar{y}_{..} } \)

Substitute \( \hat{\mu} \) into (2) and (3)

\[
\begin{align*}
(2) & \quad b n \bar{y}_{..} + b n \hat{\alpha}_i = y_{i..} \rightarrow \bar{y}_{..} + \hat{\alpha}_i = \frac{y_{i..}}{b} \rightarrow \boxed{\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{..} } \\
(3) & \quad a n \bar{y}_{..} + a n \hat{\beta}_j = y_{..j} \rightarrow \bar{y}_{..} + \hat{\beta}_j = \frac{y_{..j}}{a} \rightarrow \boxed{\hat{\beta}_j = \bar{y}_{..j} - \bar{y}_{..} } \\
\end{align*}
\]

Substitute \( \hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j \) into (4)

\[
\begin{align*}
(4) & \quad n \bar{y}_{..} + n(\bar{y}_{i..} - \bar{y}_{..}) + n(\bar{y}_{..j} - \bar{y}_{..}) + n(\hat{\alpha}\hat{\beta})_{ij} = y_{ij} \\
& \quad \bar{y}_{..} + \bar{y}_{i..} - \bar{y}_{..} + \bar{y}_{..j} - \bar{y}_{..} + (\hat{\alpha}\hat{\beta})_{ij} = \frac{y_{ij}}{n} \\
\end{align*}
\]

\[ \boxed{(\hat{\alpha}\hat{\beta})_{ij} = \frac{y_{ij} - y_{i..} - y_{..j} + y_{..}}{n} } \]

\[
\begin{array}{|c|c|c|}
\hline
\hat{\alpha}_{11} & \hat{\alpha}_{12} & \ldots & \hat{\alpha}_{1b} \\
\hline
\hat{\alpha}_{21} & \hat{\alpha}_{22} & \ldots & \hat{\alpha}_{2b} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\alpha}_{a1} & \hat{\alpha}_{a2} & \ldots & \hat{\alpha}_{ab} \\
\hline
\end{array}
\]

\[
\begin{align*}
\sum_{i=1}^{a} \sum_{j=1}^{b} (\hat{\alpha}\hat{\beta})_{ij} = 0 \\
\rightarrow (a-1)(b-1) \text{ Linearly Independent Elements} \\
\rightarrow (a-1)(b-1) \text{ Linearly Dependent Elements (Border)} \\
\rightarrow (a-1)(b-1) \text{ d.F. For Interaction}
\end{align*}
\]
4.3 Matrix Forms for the Twoway ANOVA

Example: Consider a completely randomized $2 \times 3$ factorial design with $n = 2$ replications for each of the six combinations of the two factors ($A$ and $B$). The following table summarizes the results:

<table>
<thead>
<tr>
<th>Factor $A$ Levels</th>
<th>Factor $B$ Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 2</td>
</tr>
<tr>
<td>2</td>
<td>3, 5</td>
</tr>
</tbody>
</table>

- Model: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$ for $i = 1, 2$ $j = 1, 2, 3$ $k = 1, 2$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$
- Assume (i) $\sum_{i=1}^{2} \alpha_i = 0$ (ii) $\sum_{j=1}^{3} \beta_j = 0$
- Thus, for the main effect constraints, we have $\alpha_2 = -\alpha_1$ and $\beta_3 = -\beta_1 - \beta_2$.
- The interaction effect constraints can be written in terms of just $\alpha\beta_{11}$ and $\alpha\beta_{12}$:

$$\alpha\beta_{12} = \alpha\beta_{22} = \alpha\beta_{13} = \alpha\beta_{23} =$$

- Thus, the reduced form of model matrix $X$ requires only 6 columns: $\mu, \alpha_1, \beta_1, \beta_2, \alpha\beta_{11}$ and $\alpha\beta_{12}$.

\[
X = \begin{bmatrix}
\mu & \alpha_1 & \beta_1 & \beta_2 & \alpha\beta_{11} & \alpha\beta_{12} \\
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 0 & 1 & 0 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1 \\
\end{bmatrix}, \quad y = \begin{bmatrix}
1 \\
2 \\
4 \\
6 \\
5 \\
6 \\
3 \\
5 \\
7 \\
4 \\
6 \\
\end{bmatrix}
\]

\[
X'X = \begin{bmatrix}
12 & 0 & 0 & 0 & 0 & 0 \\
0 & 12 & 0 & 0 & 0 & 0 \\
0 & 0 & 8 & 4 & 0 & 0 \\
0 & 0 & 4 & 8 & 0 & 0 \\
0 & 0 & 0 & 8 & 4 & 0 \\
0 & 0 & 0 & 0 & 4 & 8 \\
\end{bmatrix}, \quad X'y = \begin{bmatrix}
54 \\
-6 \\
-10 \\
1 \\
-6 \\
-3 \\
\end{bmatrix}
\]

\[
(X'X)^{-1} = \frac{1}{12} \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & -1 & 0 & 0 \\
0 & 0 & -1 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & -1 \\
0 & 0 & 0 & 0 & -1 & 2 \\
\end{bmatrix}, \quad (X'X)^{-1}X'y = \begin{bmatrix}
4.5 \\
-0.5 \\
-1.75 \\
1 \\
-0.75 \\
0 \\
\end{bmatrix}
\]

Thus, $\hat{\alpha}_2 = -\hat{\alpha}_1 = 0.5$ $\hat{\beta}_3 = -\hat{\beta}_1 - \hat{\beta}_2 = 0.75$ $\hat{\alpha}_2\beta_{11} = -\hat{\alpha}_1\beta_{11} = 0.75$ $\hat{\alpha}_2\beta_{12} = -\hat{\alpha}_1\beta_{12} = 0.75$ $\hat{\beta}_2 = \hat{\alpha}_1\beta_{11} = 0.75$ $\hat{\alpha}_2\beta_{12} = \hat{\alpha}_1\beta_{12} = -0.75$
Alternate Approach: Keeping $1 + a + b + (a \ast b)$ Columns

$$\mu \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \alpha_1 \beta_2 \quad \alpha_2 \beta_2 \quad \alpha_1 \beta_3 \quad \alpha_2 \beta_3$$

$$X = \begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}$$

$$y = \begin{bmatrix}
1 \\
2 \\
4 \\
6 \\
6 \\
3 \\
5 \\
5 \\
5 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \end{bmatrix}$$

$$X'X = \begin{bmatrix}
12 & 6 & 6 & 4 & 4 & 4 & 2 & 2 & 2 & 2 & 2 & 2 \\
6 & 7 & 1 & 2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\
6 & 1 & 7 & 2 & 2 & 2 & 0 & 0 & 2 & 2 & 2 & 2 \\
4 & 2 & 2 & 5 & 1 & 1 & 2 & 0 & 0 & 2 & 0 & 0 \\
4 & 2 & 2 & 1 & 5 & 1 & 1 & 0 & 2 & 0 & 0 & 2 \\
4 & 2 & 2 & 1 & 1 & 5 & 0 & 0 & 2 & 0 & 0 & 2 \\
2 & 2 & 0 & 2 & 0 & 0 & 4 & 1 & 1 & 1 & 1 & 0 \\
2 & 2 & 0 & 0 & 2 & 0 & 1 & 4 & 1 & 0 & 1 & 0 \\
2 & 2 & 0 & 0 & 0 & 2 & 1 & 1 & 4 & 0 & 0 & 1 \\
2 & 0 & 2 & 2 & 0 & 0 & 1 & 0 & 0 & 4 & 1 & 1 \\
2 & 0 & 2 & 0 & 2 & 0 & 0 & 1 & 0 & 1 & 4 & 1 \\
2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 1 & 1 & 1 & 4 \\
\end{bmatrix}$$

$$X'y = \begin{bmatrix}
54 \\
24 \\
30 \\
11 \\
22 \\
21 \\
3 \\
10 \\
11 \\
8 \\
12 \\
10 \end{bmatrix}$$

$$X'X^{-1} = \frac{1}{180}$$

$$(X'X)^{-1}X'y = \begin{bmatrix}
4.5 \\
-5 \\
-5 \\
-1.75 \\
1 \\
.75 \\
-7.5 \\
0 \\
.75 \\
0 \\
-7.5 \\
\end{bmatrix}$$

$$\begin{bmatrix}
\hat{\mu} \\
\alpha_1 \\
\alpha_2 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\alpha_{11} \\
\alpha_{12} \\
\alpha_{13} \\
\alpha_{21} \\
\alpha_{22} \\
\alpha_{23} \\
\end{bmatrix}$$
4.4 Notation for an ANOVA

- \( SS_A = nb \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 = \) the sum of squares for factor \( A \) (\( df = a - 1 \))

\[ MS_A = SS_A / (a - 1) \]

- \( MS_A = SS_A / (a - 1) = \) the mean square for factor \( A \)

- \( SS_B = na \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 = \) the sum of squares for factor \( B \) (\( df = b - 1 \))

\[ MS_B = SS_B / (b - 1) \]

- \( MS_B = SS_B / (b - 1) = \) the mean square for factor \( B \)

- \( SS_{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} ([\bar{y}_{ij} - \bar{y}_{.j} - (\bar{y}_{i.} - \bar{y}_{..})]^2 = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 = \) the \( A \times B \) interaction sum of squares (\( df = (a - 1)(b - 1) \))

\[ MS_{AB} = SS_{AB} / (a - 1)(b - 1) = \] the mean square for the \( A \times B \) interaction

- \( SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_i} (y_{ijk} - \bar{y}_{ijk})^2 = \) the error sum of squares (\( df = ab(n - 1) \))

\[ MS_E = SS_E / ab(n - 1) = \] the mean square error

- \( SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{..})^2 = \) the total sum of squares (\( df = abn - 1 \))

The total sum of squares is partitioned into components corresponding to the terms in the model:

\[ \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{..})^2 = nb \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 + na \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 + n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i.})^2 \]

OR

- The alternate \( SS \) formulas for the balanced two factorial design are:

\[ SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}^2 - y_{..}^2 / abn \]

\[ SS_A = \sum_{i=1}^{a} y_{i.}^2 - y_{..}^2 / abn \]

\[ SS_B = \sum_{j=1}^{b} y_{.j}^2 - y_{..}^2 / abn \]

\[ SS_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij}^2 / n \]

\[ SS_E = SS_T - SS_A - SS_B - SS_{AB} \]

- The alternate \( SS \) formulas for the unbalanced two factorial design are:

\[ SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}^2 / N \]

\[ SS_A = \sum_{i=1}^{a} y_{i.}^2 / n_i - y_{..}^2 / N \]

\[ SS_B = \sum_{j=1}^{b} y_{.j}^2 / n_j - y_{..}^2 / N \]

\[ SS_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij}^2 / n_{ij} - SS_A - SS_B - y_{..}^2 / N \]

\[ SS_E = SS_T - SS_A - SS_B - SS_{AB} \]

where \( N = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} \), \( n_i = \sum_{j=1}^{b} n_{ij} \), \( n_j = \sum_{i=1}^{a} n_{ij} \).
### Balanced Two-Factor Factorial ANOVA Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>SS_A</td>
<td>a − 1</td>
<td>MS_A = SS_A/(a − 1)</td>
<td>F_A = MS_A/MS_E</td>
</tr>
<tr>
<td>B</td>
<td>SS_B</td>
<td>b − 1</td>
<td>MS_B = SS_B/(b − 1)</td>
<td>F_B = MS_B/MS_E</td>
</tr>
<tr>
<td>A * B</td>
<td>SS_AB</td>
<td>(a − 1)(b − 1)</td>
<td>MS_AB = SS_AB/(a − 1)(b − 1)</td>
<td>F_A*B = MS_AB/MS_E</td>
</tr>
<tr>
<td>Error</td>
<td>SS_E</td>
<td>ab(n − 1)</td>
<td>MS_E = SS_E/(ab(n − 1))</td>
<td>——</td>
</tr>
<tr>
<td>Total</td>
<td>SS_total</td>
<td>abn − 1</td>
<td>——</td>
<td>——</td>
</tr>
</tbody>
</table>

For the unbalanced case, replace ab(n − 1) with N − ab for the d.f. for SS_E and replace abn − 1 with N − 1 for the d.f. for SS_total where \( N = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} \).

#### 4.5 Comments on Interpreting the ANOVA

- **Test** \( H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0 \) vs. \( H_1 : \text{at least one } (\alpha\beta)_{ij} \neq (\alpha\beta)_{i'j'} \) first.
  - If this test indicates that there is not a significant interaction, then continue testing the hypotheses for the two main effects:
    \[
    H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a \quad \text{vs.} \quad H_1 : \text{at least one } \alpha_i \neq \alpha_i' \\
    H_0 : \beta_1 = \beta_2 = \cdots = \beta_b \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq \beta_j' 
    \]
  - If this test indicates that there is a significant interaction, then the interpretation of significant main effects hypotheses can be masked. To draw conclusions about a main effect, we will fix the levels of one factor and vary the levels of the other. Using this approach (combined with interaction plots) we may be able to provide an interpretation of main effects.
- If we assume the constraints in [24], then the hypotheses can be rewritten as:
  \[
  H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0 \quad \text{vs.} \quad H_1 : \text{at least one } (\alpha\beta)_{ij} \neq 0 \\
  H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \alpha_i \neq 0 \\
  H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0 
  \]

#### 4.6 ANOVA for a 2 × 2 Factorial Design Example

- We will now use SAS to analyze the 2 × 2 factorial design data discussed earlier.

<table>
<thead>
<tr>
<th>M</th>
<th>Medium 1</th>
<th>Medium 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>hours</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>21 23 20</td>
<td>25 24 29</td>
</tr>
<tr>
<td>22 28 26</td>
<td>26 25 27</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>37 38 35</td>
<td>31 29 30</td>
</tr>
<tr>
<td>39 38 36</td>
<td>34 33 35</td>
<td></td>
</tr>
</tbody>
</table>
### ANOVA and Estimation of Effects for a 2x2 Design

**The GLM Procedure**

**Dependent Variable: growth**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>691.45833333</td>
<td>230.4861111</td>
<td>45.12</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>102.16666667</td>
<td>5.10833333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>23</td>
<td>793.62500000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>growth Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.871266</td>
<td>7.629240</td>
<td>2.260162</td>
<td>29.625000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>590.0416667</td>
<td>590.0416667</td>
<td>115.51</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>medium</td>
<td>1</td>
<td>9.3750000</td>
<td>9.3750000</td>
<td>1.84</td>
<td>0.1906</td>
</tr>
<tr>
<td>time*medium</td>
<td>1</td>
<td>92.0416667</td>
<td>92.0416667</td>
<td>18.02</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

| Parameter     | Estimate   | Standard Error | t Value | Pr > |t| |
|---------------|------------|----------------|---------|------|--|
| Intercept     | 32.00000000 | 0.92270737 | 34.68   | <.0001 |
| time 12       | -6.00000000 | 1.30490528 | -4.60   | 0.0002 |
| time 18       | 0.00000000  | .             | .       | .    |
| medium 1      | 5.16666667  | 1.30490528 | 3.96    | 0.0008 |
| medium 2      | 0.00000000  | .             | .       | .    |
| time*medium 12 1 | -7.83333333 | 1.84541474 | -4.24   | 0.0004 |
| time*medium 12 2 | 0.00000000 | .             | .       | .    |
| time*medium 18 1 | 0.00000000 | .             | .       | .    |
| time*medium 18 2 | 0.00000000 | .             | .       | .    |

**Note:** The XX matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

| Parameter     | Estimate   | Standard Error | t Value | Pr > |t| |
|---------------|------------|----------------|---------|------|--|
| mu            | 29.6250000 | 0.46135368     | 64.21   | <.0001 |
| time=12       | -4.9583333 | 0.46135368     | -10.75  | <.0001 |
| time=18       | 4.9583333  | 0.46135368     | 10.75   | <.0001 |
| medium=1      | 0.6250000  | 0.46135368     | 1.35    | 0.1906 |
| medium=2      | -0.6250000 | 0.46135368     | -1.35   | 0.1906 |
| time=12 medium=1 | -1.9583333 | 0.46135368     | -4.24   | 0.0004 |
| time=12 medium=2 | 1.9583333  | 0.46135368     | 4.24    | 0.0004 |
| time=18 medium=1 | 1.9583333  | 0.46135368     | 4.24    | 0.0004 |
| time=18 medium=2 | -1.9583333 | 0.46135368     | -4.24   | 0.0004 |
4.6.1 SAS Code for 2 x 2 Factorial Design

DM 'LOG; CLEAR; OUT; CLEAR;';

ODS GRAPHICS ON;
ODS PRINTER PDF file='C:\COURSES\ST541\TWOWAY1.PDF';
OPTIONS NODATE NONUMBER;

************************************************;
*** EXAMPLE: 2-FACTOR FACTORIAL (2x2) DESIGN ***;
************************************************;
DATA in;
  DO time = 12 to 18 by 6;
  DO medium = 1 to 2;
  DO rep = 1 to 6;
    INPUT growth @@; OUTPUT;
  END; END; END;
CARDS;
21 23 20 22 28 26 25 24 29 26 25 27
37 38 35 39 38 36 31 29 30 34 33 35
; PROC GLM DATA=in PLOTS=(ALL);
  CLASS time medium;
  MODEL growth = time|medium / SS3 SOLUTION;
  MEANS time|medium;

*** Estimate mu ***;
  ESTIMATE 'mu' intercept 1;

*** Estimate the main effects for factor time';
  ESTIMATE 'time=12' time 1 -1 / divisor = 2 ;
  ESTIMATE 'time=18' time -1 1 / divisor = 2 ;

*** Estimate the main effects for factor medium';
  ESTIMATE 'medium=1' medium 1 -1 / divisor = 2 ;
  ESTIMATE 'medium=2' medium -1 1 / divisor = 2 ;

*** Estimate the interaction effects';
  *** Take the product of the tau_i and beta_j coefficients;
  *** from the main effects ESTIMATE statement. Divisor = a*b;

  *** To estimate taubeta i,j
  *** (1 -1) x (1 -1) = (1 -1 -1 1) for i,j = 12,1;
  *** (1 -1) x (-1 1) = (-1 1 1 -1) for i,j = 12,2;
  *** (-1 1) x (1 -1) = (-1 1 1 -1) for i,j = 18,1;
  *** (-1 1) x (-1 1) = (1 -1 -1 1) for i,j = 18,2;

  ESTIMATE 'time=12 medium=1' time*medium 1 -1 -1 1 / divisor = 4;
  ESTIMATE 'time=12 medium=2' time*medium -1 1 1 -1 / divisor = 4;
  ESTIMATE 'time=18 medium=1' time*medium -1 1 1 -1 / divisor = 4;
  ESTIMATE 'time=18 medium=2' time*medium 1 -1 -1 1 / divisor = 4;

OUTPUT OUT=diag P=pred R=resid;

TITLE 'ANOVA and Estimation of Effects for a 2x2 Design';
PROC UNIVARIATE DATA=diag NORMAL;
  VAR resid;
RUN;

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W</td>
<td>0.966156</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D</td>
<td>0.140751</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq</td>
<td>0.049547</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq</td>
<td>0.303234</td>
</tr>
</tbody>
</table>
4.7 Tests of Normality (Supplemental)

- For an ANOVA, we assume the errors are normally distributed with mean 0 and constant variance \( \sigma^2 \). That is, we assume the random error \( \epsilon \sim N(0, \sigma^2) \).

- The Kolmogorov-Smirnov Goodness-of-Fit Test, the Cramer-Von Mises Goodness-of-Fit Test, and the Anderson-Darling Goodness-of-Fit Test can be applied to any distribution \( F(x) \).

- Although the following notes use the general form \( F(x) \), we will be assuming \( F(x) \) represents a normal distribution with mean 0 and constant variance.

- We are also assuming that the random sample referred to in each test is the set of residuals from the ANOVA.

- Thus, in each test we are checking the normality assumption in the ANOVA. In this case, we want to see a large \( p \)-value because we do not want to reject the null hypothesis that the errors are normally distributed.

4.7.1 Kolmogorov-Smirnov Goodness-of-Fit Test

**Assumptions:** Given a random sample of \( n \) independent observations

- The measurement scale is at least ordinal.

- The observations are sampled from a continuous distribution \( F(x) \).

**Hypotheses:** For a hypothesized distribution \( F^*(x) \)

(i) Two-sided: \( H_0 : F(x) = F^*(x) \) for all \( x \) vs. \( H_1 : F(x) \neq F^*(x) \) for some \( x \)

(ii) One-sided: \( H_0 : F(x) \geq F^*(x) \) for all \( x \) vs. \( H_1 : F(x) < F^*(x) \) for some \( x \)

(iii) One-sided: \( H_0 : F(x) \leq F^*(x) \) for all \( x \) vs. \( H_1 : F(x) > F^*(x) \) for some \( x \)

**Method:** For a given \( \alpha \)

- Define the empirical distribution function \( S_n(x) = \frac{\text{Number of observations} \leq x}{n} \)

(i) Two-sided test statistic: \( T = \sup_x |F^*(x) - S_n(x)| \)

- When plotted, \( T \) is the greatest vertical difference between the empirical and the hypothesized distribution.

(ii) One-sided test statistic: \( T^+ = \sup_x (F^*(x) - S_n(x)) \)

(iii) One-sided test statistic: \( T^- = \sup_x (S_n(x) - F^*(x)) \)

**Decision Rule**

- Critical values for \( T, T^+ \) and \( T^- \) are found in nonparametrics textbooks. For larger samples sizes, an asymptotic critical value can be used.

- We will just rely on \( p \)-values to make a decision.
4.7.2 Cramer-Von Mises Goodness-of-Fit Test

**Assumptions:** Same as the Kolmogorov-Smirnov test

**Hypotheses:** For a hypothesized distribution \( F^*(x) \)

\[ H_0 : F(x) = F^*(x) \quad \text{for all } x \quad \text{vs.} \quad H_1 : F(x) \neq F^*(x) \quad \text{for some } x \]

**Method:** For a given \( \alpha \)

- Define the empirical distribution function \( S_n(x) = \frac{\text{Number of observations } \leq x}{n} \)
- The Cramer-von Mises test statistic \( W^2 \) is defined to be

\[ W^2 = n \int_{-\infty}^{\infty} [F^*(x) - S_n(x)]^2 dF^*(x). \]

- This form can reduces to

\[ W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left( F^*(x(i)) - \frac{2i-1}{2n} \right)^2 \]

where \( x(1), x(2), \ldots, x(n) \) represents the ordered sample in ascending order.

**Decision Rule**

- Tables of critical values exist for the exact distribution of \( W^2 \) when \( H_0 \) is true. Computers generate critical values for the asymptotic \((n \to \infty)\) distribution of \( W^2 \).
- If \( W^2 \) becomes too large (or \( p\)-value < \( \alpha \)), then we will **Reject** \( H_0 \).

4.7.3 Anderson-Darling Goodness-of-Fit Test

**Assumptions:** Same as the Kolmogorov-Smirnov and Cramer-von Mises tests

**Hypotheses:** Same as the Cramer-von Mises test.

**Method:** For a given \( \alpha \)

- Define the empirical distribution function \( S_n(x) = \frac{\text{Number of observations } \leq x}{n} \)
- The Anderson-Darling test statistic \( A^2 \) is defined to be

\[ A^2 = \int_{-\infty}^{\infty} \frac{1}{F^*(x)(1-F^*(x))} (F^*(x) - S_n(x))^2 \, dx. \]

- This form can reduces to

\[ A^2 = -\frac{1}{n} \left\{ (2i-1) \left( \ln F^*(x(i)) + \ln(1 - F^*(x(n+1-i))) \right) \right\} - n \]

where \( x(1), x(2), \ldots, x(n) \) represents the ordered sample in ascending order.

**Decision Rule**

- Computers generate critical values for the asymptotic \((n \to \infty)\) distribution of \( A^2 \).
- If \( A^2 \) becomes too large (or \( p\)-value < \( \alpha \)), then we will **Reject** \( H_0 \).