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.

TYPE	TOTALS	MEANS	(if $n_{ij} = n$)
Cell(i,j)	$y_{ij\cdot} = \sum_{k=1}^{n_{ij}} y_{ijk}$	$\overline{y}_{ij\cdot} = y_{ij\cdot}/n_{ij}$	$= y_{ij\cdot}/n$
i^{th} level of A	$y_{i\cdot\cdot} = \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}$	$\overline{y}_{i\cdot\cdot} = y_{i\cdot\cdot}/\sum_{j=1}^{b} n_{ij}$	$= y_{i\cdot\cdot}/bn$
j^{th} level of B	$y_{\cdot j\cdot} = \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} y_{ijk}$	$\overline{y}_{\cdot j\cdot} = y_{\cdot j\cdot}/\sum_{i=1}^{a} n_{ij}$	$= y_{\cdot j\cdot}/an$
Overall	$y_{\cdots} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}$	$\overline{y}_{\cdots} = y_{\cdots}/\sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}$	$= y_{\cdots}/abn$

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

EXAMPLE (A 2×2 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 $M * T$ treatment combinations. The $N = 24$ measurements were taken in a completely randomized order. The results:

THE DATA

M

		Medium 1	Medium 2
12		21 23 20	25 24 29
T	hours	22 28 26	26 25 27
18		37 38 35	31 29 30
hours		39 38 36	34 33 35

i = Level of T j = Level of M
k = Observation number
$y_{ijk} = k^{th}$ observation from the ith level of T and j^{th} level of M

TOTALS

	$T = 1$	$T = 2$	
$T = 12$	$y_{11\cdot} = 140$	$y_{12\cdot} = 156$	$y_{1\cdot\cdot} = 296$
$T = 18$	$y_{21\cdot} = 223$	$y_{22\cdot} = 192$	$y_{2\cdot\cdot} = 415$
	$y_{\cdot 1\cdot} = 363$	$y_{\cdot 2\cdot} = 348$	$y_{\cdots} = 711$

MEANS

	$M = 1$	$M = 2$	
$T = 12$	$\overline{y}_{11\cdot} = 23.\overline{3}$	$\overline{y}_{12\cdot} = 26$	$\overline{y}_{1\cdot\cdot} = 24.\overline{6}$
$T = 18$	$\overline{y}_{21\cdot} = 37.1\overline{6}$	$\overline{y}_{22\cdot} = 32$	$\overline{y}_{2\cdot\cdot} = 34.58\overline{3}$
	$\overline{y}_{\cdot 1\cdot} = 30.25$	$\overline{y}_{\cdot 2\cdot} = 29.00$	$\overline{y}_{\cdots} = 29.625$

- The effect of changing T from 12 to 18 hours on the response depends on the level of M.

    - For medium 1, the T effect $=$ $37.1\overline{6} - 23.\overline{3}$ $=$
    - 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 $=$ $26 - 23.\overline{3}$ $=$
    - For $T = 18$ hours, the M effect $=$ $32 - 37.1\overline{6}$ $=$

- 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 2×2 design example:

  - If $13.8\overline{3}$ is significantly different than 6 for the M effects, then we have a significant $M * T$ interaction.

    Or,

  - If $2.\overline{6}$ is significantly different than $-5.1\overline{6}$ for the T effects, then we have a significant $M * T$ 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 smal values of the MS_E, even small interaction effects (less nonparallelism) may be significant.

- When an $A * B$ interaction is large, the corresponding main effects A and B may have little practical meaning. Knowledge of the $A * B$ 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:

$$y_{ijk} = \tag{22}$$

where μ is the baseline mean, α_i is the ith factor A effect, β_j is the jth factor B effect, $(\alpha\beta)_{ij}$ is the $(i,j)^{th}$ $A*B$ interaction effect, ϵ_{ijk} is the random error of the kth 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 <u>fixed</u>.

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

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \tag{23}$$

- If we impose the constraints

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0 \qquad \sum_{i=1}^{a}(\alpha\beta)_{ij} = 0 \ \text{ for } \ \text{all } j \qquad \text{and} \qquad \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0 \ \text{ for } \ \text{all } i, \tag{24}$$

then the least squares estimates of the model parameters are

$$\widehat{\mu} = \qquad\qquad \widehat{\alpha}_i = \qquad\qquad \widehat{\beta}_j =$$

$$\widehat{\alpha\beta}_{ij} =$$

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

$$y_{ijk} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\alpha\beta}_{ij} + e_{ijk}$$

$$= \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + e_{ijk}$$

where e_{ijk} is the kth residual from the treatment $(i,j)^{th}$ cell, and $e_{ijk} =$

- For the 2×2 design,

$$\bar{y}_{...} = 29.625 \qquad \bar{y}_{1..} = 24.\overline{6} \qquad \bar{y}_{2..} = 34.58\overline{6} \qquad \bar{y}_{.1} = 30.25 \qquad \bar{y}_{.2} = 29.00$$

- Assuming the constraints in (24),

$$\alpha_1 = 24.\overline{6} - 29.625 =$$
$$\alpha_2 = 34.58\overline{3} - 29.625 =$$
$$\beta_1 = 30.25\overline{6} - 29.625 =$$
$$\beta_2 = 29.00\overline{6} - 29.625 =$$
$$\alpha\beta_{11} = 23.\overline{3} - 24.\overline{6} - 30.25 + 29.625 =$$
$$\alpha\beta_{12} = 26 - 24.\overline{6} - 29.00 + 29.625 =$$
$$\alpha\beta_{21} = 37.1\overline{6} - 34.58\overline{3} - 30.25 + 29.625 =$$
$$\alpha\beta_{22} = 32 - 34.58\overline{3} - 29.00 + 29.625 =$$

ESTIMATION OF PARAMETERS   MODEL: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

GOAL: MINIMIZE $L = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} \sum\limits_{k=1}^{n} \left(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - (\widehat{\alpha\beta})_{ij} \right)^2$

(1)	(3)				(2)	
(1) $\hat{\mu}$	$abn\hat{\mu}$	$+ bn \sum\limits_i \hat{\alpha}_i$	$+ an \sum\limits_j \hat{\beta}_j$	$+ n \sum\limits_i \sum\limits_j (\widehat{\alpha\beta})_{ij}$	$= y_{...}$	
(2) $\hat{\alpha}_i$	$bn\hat{\mu}$	$+ bn\hat{\alpha}_i$	$+ n \sum\limits_j \hat{\beta}_j$	$+ n \sum\limits_j (\widehat{\alpha\beta})_{ij}$	$= y_{i..}$	$i = 1, \ldots, a$
(3) $\hat{\beta}_j$	$an\hat{\mu}$	$+ n \sum\limits_i \hat{\alpha}_i$	$+ an \hat{\beta}_j$	$+ n \sum\limits_i (\widehat{\alpha\beta})_{ij}$	$= y_{.j.}$	$j = 1, \ldots, b$
(4) $(\widehat{\alpha\beta})_{ij}$	$n\hat{\mu}$	$+ n\hat{\alpha}_i$	$+ n\hat{\beta}_j$	$+ n\widehat{\alpha\beta}_{ij}$	$= y_{ij.}$	

CONSTRAINTS: $\sum\limits_i \hat{\alpha}_i = 0 \qquad \sum\limits_j \hat{\beta}_j = 0 \qquad \sum\limits_i (\widehat{\alpha\beta})_{ij} = 0 \qquad \sum\limits_j (\widehat{\alpha\beta})_{ij} = 0$

FROM (1) $\qquad abn\hat{\mu} = y_{...} \quad \rightarrow \quad \boxed{\hat{\mu} = \overline{y}_{...}}$

SUBSTITUTE $\hat{\mu}$ INTO (2) AND (3)

$\rightarrow$ (2) $\quad bn\overline{y}_{...} + bn\hat{\alpha}_i = y_{i..} \quad \rightarrow \quad \overline{y}_{...} + \hat{\alpha}_i = \overline{y}_{i..} \quad \rightarrow \boxed{\hat{\alpha}_i = \overline{y}_{i..} - \overline{y}_{...}}$

$\rightarrow$ (3) $\quad an\overline{y}_{...} + an\hat{\beta}_j = y_{.j.} \quad \rightarrow \quad \overline{y}_{...} + \hat{\beta}_j = \overline{y}_{.j.} \quad \rightarrow \boxed{\hat{\beta}_j = \overline{y}_{.j.} - \overline{y}_{...}}$

SUBSTITUTE $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j$ INTO (4)

$\rightarrow$ (4) $\quad n\overline{y}_{...} + n(\overline{y}_{i..} - \overline{y}_{...}) + n(\overline{y}_{.j.} - \overline{y}_{...}) + n(\widehat{\alpha\beta})_{ij} = y_{ij.}$

$\overline{y}_{...} + \overline{y}_{i..} - \overline{y}_{...} + \overline{y}_{.j.} - \overline{y}_{...} + (\widehat{\alpha\beta})_{ij} = \overline{y}_{ij.}$

$\boxed{(\widehat{\alpha\beta})_{ij} = \overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}}$

$\widehat{\alpha\beta}_{11}$	$\widehat{\alpha\beta}_{12}$	$\cdots$	$\cdots$	$\widehat{\alpha\beta}_{1b}$
$\widehat{\alpha\beta}_{21}$	$\widehat{\alpha\beta}_{22}$	$\cdots$	$\cdots$	$\widehat{\alpha\beta}_{2b}$
$\vdots$	$\vdots$			$\vdots$
$\widehat{\alpha\beta}_{a1}$	$\widehat{\alpha\beta}_{a2}$	$\cdots$	$\cdots$	$\widehat{\alpha\beta}_{ab}$

1   b-1

$\sum\limits_i (\widehat{\alpha\beta})_{ij} = \sum\limits_j (\widehat{\alpha\beta})_{ij} = 0$

$\rightarrow (a-1)(b-1)$ LINEARLY INDEPENDENT ELEMENTS

$\rightarrow a+b-1$ LINEARLY DEPENDENT ELEMENTS (BORDER)

$\rightarrow (a-1)(b-1)$ d.F. FOR INTERACTION

4.3 Matrix Forms for the Twoway ANOVA

Example: Consider a completely randomized 2×3 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:

Factor A	Factor B Levels		
Levels	1	2	3
1	1 , 2	4 , 6	5 , 6
2	3 , 5	5 , 7	4 , 6

- 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) $\sum_{j=1}^{3} \beta_j = 0$

  (iii) $\sum_{j=1}^{3} (\alpha\beta)_{ij} = 0$ for $i = 1, 2$ (iv) $\sum_{i=1}^{2} (\alpha\beta)_{ij} = 0$ for $j = 1, 2, 3$

- 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 $\alpha\beta_{12}$:

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

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

$$
\begin{array}{cccccc}
\mu & \alpha_1 & \beta_1 & \beta_2 & \alpha\beta_{11} & \alpha\beta_{12}
\end{array}
$$

$$
X = \begin{bmatrix}
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 0 & 1 & 0 & -1 \\
1 & -1 & 0 & 1 & 0 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1
\end{bmatrix}
\qquad
y = \begin{bmatrix}
1 \\ 2 \\ 4 \\ 6 \\ 5 \\ 6 \\ 3 \\ 5 \\ 5 \\ 7 \\ 4 \\ 6
\end{bmatrix}
$$

$$
X'X = \begin{bmatrix}
12 & 0 & 0 & 0 & 0 & 0 \\
0 & 12 & 0 & 0 & 0 & 0 \\
0 & 0 & 8 & 4 & 0 & 0 \\
0 & 0 & 4 & 8 & 0 & 0 \\
0 & 0 & 0 & 0 & 8 & 4 \\
0 & 0 & 0 & 0 & 4 & 8
\end{bmatrix}
\qquad
X'y = \begin{bmatrix}
54 \\ -6 \\ -10 \\ 1 \\ -6 \\ -3
\end{bmatrix}
$$

$$
(X'X)^{-1} = \frac{1}{12}\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & -1 & 0 & 0 \\
0 & 0 & -1 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & -1 \\
0 & 0 & 0 & 0 & -1 & 2
\end{bmatrix}
\qquad
(X'X)^{-1}X'y = \begin{bmatrix}
4.5 \\ -0.5 \\ -1.75 \\ 1 \\ -0.75 \\ 0
\end{bmatrix}
\begin{bmatrix}
\widehat{\mu} \\ \widehat{\alpha}_1 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\alpha\beta}_{11} \\ \widehat{\alpha\beta}_{12}
\end{bmatrix}
$$

Thus, $\widehat{\alpha}_2 = -\widehat{\alpha}_1 = 0.5$ $\widehat{\beta}_3 = -\widehat{\beta}_1 - \widehat{\beta}_2 = 0.75$ $\widehat{\alpha\beta}_{21} = -\widehat{\alpha\beta}_{11} = 0.75$

$\widehat{\alpha\beta}_{22} = -\widehat{\alpha\beta}_{12} = 0$ $\widehat{\alpha\beta}_{13} = -\widehat{\alpha\beta}_{11} - \widehat{\alpha\beta}_{12} = 0.75$ $\widehat{\alpha\beta}_{23} = \widehat{\alpha\beta}_{11} + \widehat{\alpha\beta}_{12} = -0.75$

Alternate Approach: Keeping $1 + a + b + (a*b)$ Columns

$$
\begin{array}{cccccccccccc}
\mu & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & \beta_3 & \alpha\beta_{11} & \alpha\beta_{12} & \alpha\beta_{13} & \alpha\beta_{21} & \alpha\beta_{22} & \alpha\beta_{23}
\end{array}
$$

$$
X = \left[\begin{array}{c|cc|ccc|cccccc}
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1
\end{array}\right]
\qquad
y = \left[\begin{array}{c}
1 \\ 2 \\ 4 \\ 6 \\ 5 \\ 6 \\ 3 \\ 5 \\ 5 \\ 7 \\ 4 \\ 6 \\ \hline
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{array}\right]
$$

$$
X'X = \left[\begin{array}{c|cc|ccc|cccccc}
12 & 6 & 6 & 4 & 4 & 4 & 2 & 2 & 2 & 2 & 2 & 2 \\ \hline
6 & 7 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 \\
6 & 1 & 7 & 2 & 2 & 2 & 0 & 0 & 0 & 2 & 2 & 2 \\ \hline
4 & 2 & 2 & 5 & 1 & 1 & 2 & 0 & 0 & 2 & 0 & 0 \\
4 & 2 & 2 & 1 & 5 & 1 & 0 & 2 & 0 & 0 & 2 & 0 \\
4 & 2 & 2 & 1 & 1 & 5 & 0 & 0 & 2 & 0 & 0 & 2 \\ \hline
2 & 2 & 0 & 2 & 0 & 0 & 4 & 1 & 1 & 1 & 0 & 0 \\
2 & 2 & 0 & 0 & 2 & 0 & 1 & 4 & 1 & 0 & 1 & 0 \\
2 & 2 & 0 & 0 & 0 & 2 & 1 & 1 & 4 & 0 & 0 & 1 \\
2 & 0 & 2 & 2 & 0 & 0 & 1 & 0 & 0 & 4 & 1 & 1 \\
2 & 0 & 2 & 0 & 2 & 0 & 0 & 1 & 0 & 1 & 4 & 1 \\
2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 1 & 1 & 1 & 4
\end{array}\right]
\qquad
X'y = \left[\begin{array}{c}
54 \\ \hline 24 \\ 30 \\ \hline 11 \\ 22 \\ 21 \\ \hline 3 \\ 10 \\ 11 \\ 8 \\ 12 \\ 10
\end{array}\right]
$$

$$
(X'X)^{-1} = \frac{1}{180}\left[\begin{array}{c|cc|ccc|cccccc}
86 & -45 & -45 & -20 & -20 & -20 & -6 & -6 & -6 & -6 & -6 & -6 \\ \hline
-45 & 70 & 20 & 0 & 0 & -0 & -10 & -10 & -10 & 10 & 10 & 10 \\
-45 & 20 & 70 & -0 & 0 & 0 & 10 & 10 & 10 & -10 & -10 & -10 \\ \hline
-20 & 0 & -0 & 80 & -10 & -10 & -30 & 15 & 15 & -30 & 15 & 15 \\
-20 & 0 & 0 & -10 & 80 & -10 & 15 & -30 & 15 & 15 & -30 & 15 \\
-20 & -0 & 0 & -10 & -10 & 80 & 15 & 15 & -30 & 15 & 15 & -30 \\ \hline
-6 & -10 & 10 & -30 & 15 & 15 & 76 & -14 & -14 & -4 & -4 & -4 \\
-6 & -10 & 10 & 15 & -30 & 15 & -14 & 76 & -14 & -4 & -4 & -4 \\
-6 & -10 & 10 & 15 & 15 & -30 & -14 & -14 & 76 & -4 & -4 & -4 \\
-6 & 10 & -10 & -30 & 15 & 15 & -4 & -4 & -4 & 76 & -14 & -14 \\
-6 & 10 & -10 & 15 & -30 & 15 & -4 & -4 & -4 & -14 & 76 & -14 \\
-6 & 10 & -10 & 15 & 15 & -30 & -4 & -4 & -4 & -14 & -14 & 76
\end{array}\right]
$$

$$
(X'X)^{-1}X'y = \left[\begin{array}{c}
4.5 \\ \hline -.5 \\ .5 \\ \hline -1.75 \\ 1 \\ .75 \\ \hline -.75 \\ 0 \\ .75 \\ .75 \\ 0 \\ -.75
\end{array}\right]
=
\left[\begin{array}{c}
\widehat{\mu} \\ \hline \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \\ \hline \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \\ \hline \widehat{\alpha\beta}_{11} \\ \widehat{\alpha\beta}_{12} \\ \widehat{\alpha\beta}_{13} \\ \widehat{\alpha\beta}_{21} \\ \widehat{\alpha\beta}_{22} \\ \widehat{\alpha\beta}_{23}
\end{array}\right]
$$

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 $(df = a - 1)$

  $MS_A = SS_A/(a - 1) =$ 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 $(df = b - 1)$

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

- $SS_{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} [(\bar{y}_{ij.} - \bar{y}_{...}) - (\bar{y}_{i..} - \bar{y}_{...}) - (\bar{y}_{.j.} - \bar{y}_{...})]^2 = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$

  $=$ the $A * B$ interaction sum of squares $(df = (a - 1)(b - 1))$

  $MS_{AB} = SS_{AB}/(a - 1)(b - 1) =$ the mean square for the $A * B$ interaction

- $SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(y_{ijk} - \bar{y}_{ij.}\right)^2 =$ 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:

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

$$+ n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

  OR

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

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn} \qquad SS_A = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{abn} \qquad SS_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{abn}$$

$$SS_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{y_{ij.}^2}{n} - SS_A - SS_B - \frac{y_{...}^2}{abn} \qquad SS_E = SS_T - SS_A - SS_B - SS_{AB}$$

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

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}^2 - \frac{y_{...}^2}{N} \qquad SS_A = \sum_{i=1}^{a} \frac{y_{i..}^2}{n_{i.}} - \frac{y_{...}^2}{N} \qquad SS_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{n_{.j}} - \frac{y_{...}^2}{N}$$

$$SS_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{y_{ij.}^2}{n_{ij}} - SS_A - SS_B - \frac{y_{...}^2}{N} \qquad SS_E = SS_T - SS_A - SS_B - SS_{AB}$$

  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

Source of Variation	Sum of Squares	d.f.	Mean Square	F Ratio
A	SS_A	$a-1$	$MS_A = SS_A/(a-1)$	$F_A = MS_A/MS_E$
B	SS_B	$b-1$	$MS_B = SS_B/(b-1)$	$F_B = MS_B/MS_E$
$A*B$	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = SS_{AB}/(a-1)(b-1)$	$F_{A*B} = MS_{AB}/MS_E$
Error	SS_E	$ab(n-1)$	$MS_E = SS_E/(ab(n-1))$	——
Total	SS_{total}	$abn-1$	——	——

For the <u>unbalanced case</u>, 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 $H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab}$ vs. H_1 : at least one $(\alpha\beta)_{ij} \neq (\alpha\beta)_{i'j'}$ first.

  - If this test indicates that there <u>is not a significant interaction</u>, 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 <u>is a significant interaction</u>, 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.

		M	
		Medium 1	Medium 2
T	12 hours	21 23 20 22 28 26	25 24 29 26 25 27
	18 hours	37 38 35 39 38 36	31 29 30 34 33 35

ANOVA and Estimation of Effects for a 2x2 Design

The GLM Procedure

Dependent Variable: growth

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	691.4583333	230.4861111	45.12	<.0001
Error	20	102.1666667	5.1083333		
Corrected Total	23	793.6250000			

R-Square	Coeff Var	Root MSE	growth Mean
0.871266	7.629240	2.260162	29.62500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
time	1	590.0416667	590.0416667	115.51	<.0001
medium	1	9.3750000	9.3750000	1.84	0.1906
time*medium	1	92.0416667	92.0416667	18.02	0.0004

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

Note: The X'X 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

Fit Diagnostics for growth
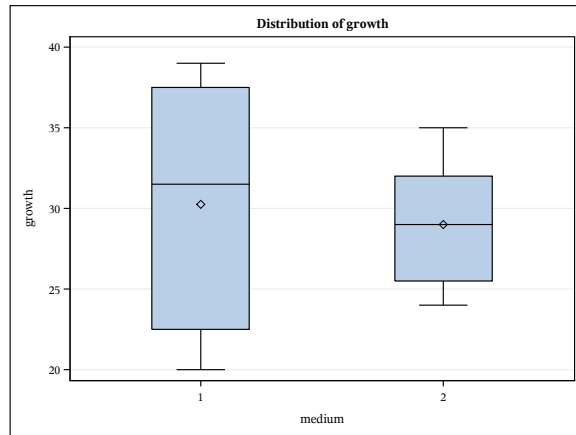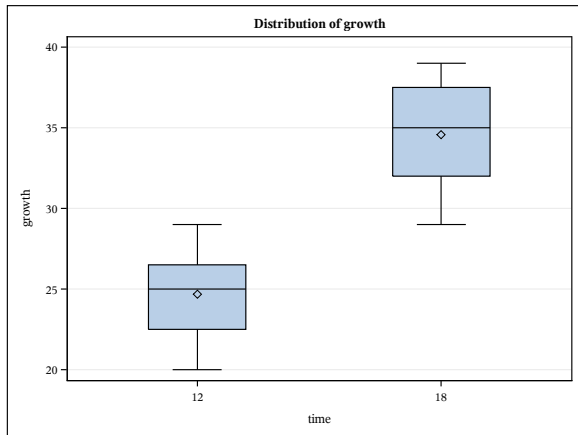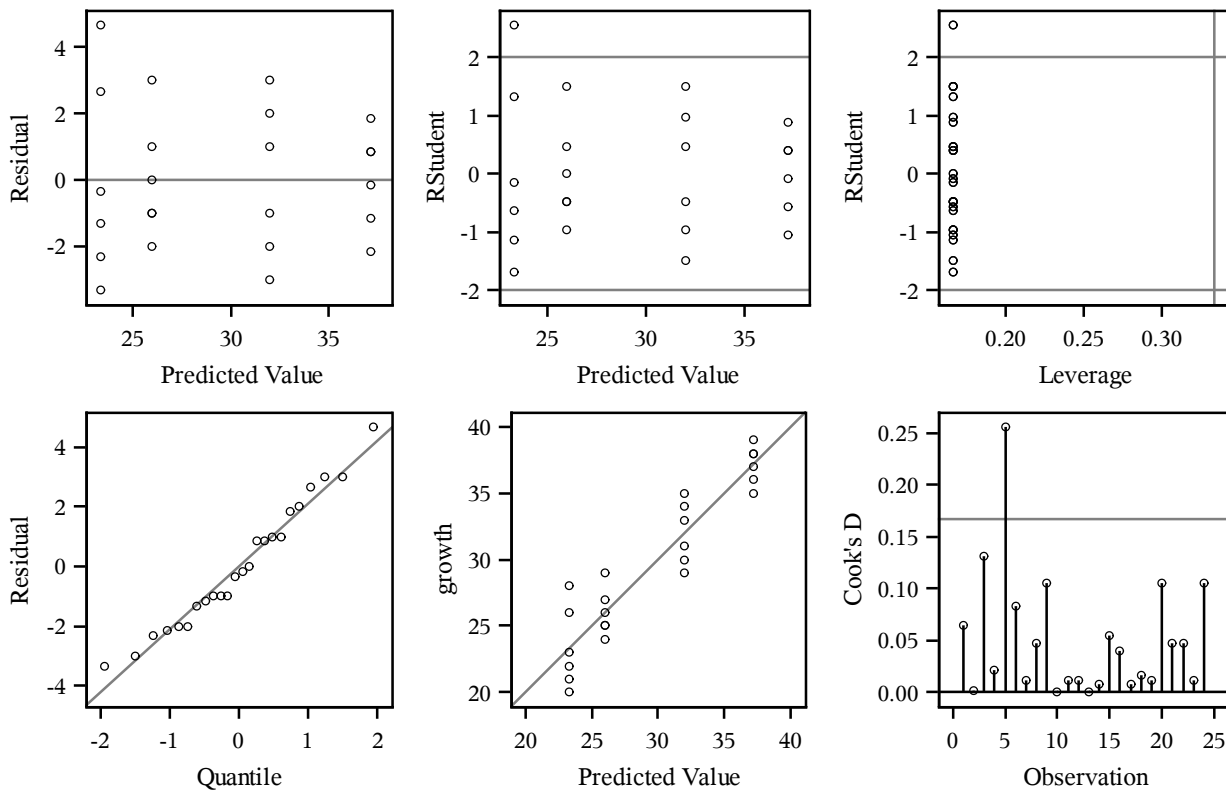
Level of time	N	growth	
		Mean	**Std Dev**
12	12	24.6666667	2.77434131
18	12	34.5833333	3.28794861

Level of medium	N	growth	
		Mean	**Std Dev**
1	12	30.2500000	7.58137670
2	12	29.0000000	3.71728151

Interaction Plot for growth

Distribution of growth
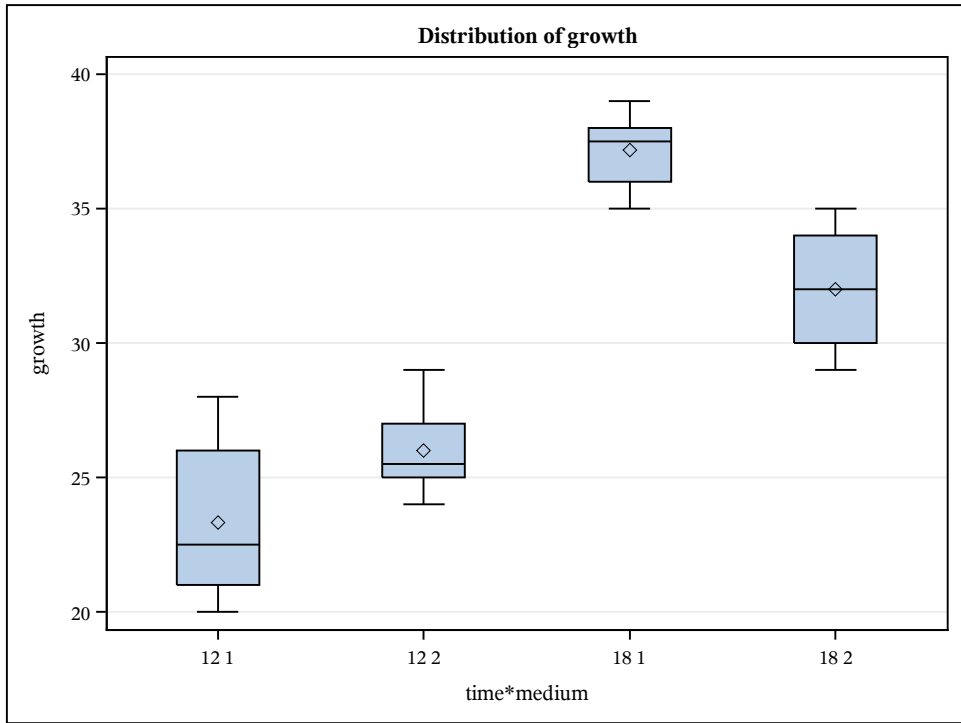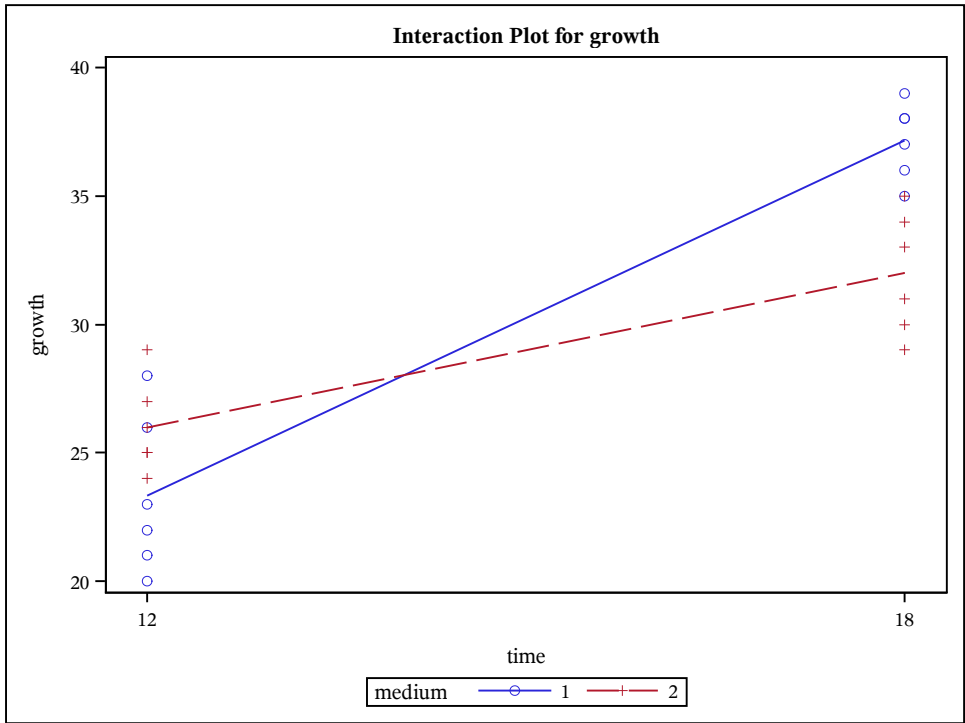
| Level of time | Level of medium | N | growth | |
			Mean	Std Dev
12	1	6	23.3333333	3.07679487
12	2	6	26.0000000	1.78885438
18	1	6	37.1666667	1.47196014
18	2	6	32.0000000	2.36643191

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.966156	Pr < W	0.5737
Kolmogorov-Smirnov	D	0.140751	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.049547	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.303234	Pr > A-Sq	>0.2500

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;
```

4.7 Tests of Normality (Supplemental)

- For an ANOVA, we assume the errors are normally distributed with mean 0 and constant variance σ^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 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 <u>vs.</u> $H_1 : F(x) \neq F^*(x)$ for some x

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

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

Method: For a given α

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

(i) Two-sided test statistic: $T = \sup\limits_{x} |F^*(x) - S_n(x)|$

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

(ii) One-sided test statistic: $T^+ = \sup\limits_{x}(F^*(x) - S_n(x))$

(iii) One-sided test statistic: $T^- = \sup\limits_{x}(S_n(x) - F^*(x))$

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 $F^*(x)$

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

Method: For a given α

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

- 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 $\quad W^2 \ = \ \dfrac{1}{12n} + \sum_{i=1}^{n} \left(F^*(x_{(i)}) - \dfrac{2i-1}{2n} \right)^2$

  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 α

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

- 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 $\quad A^2 \ = \ -\dfrac{1}{n} \left\{ (2i-1) \left(\ln F^*(x_{(i)}) + \ln(1 - F^*(x_{(n+1-i)})) \right) \right\} - n$ 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.