and the expectation of SST requires the substitution of $\overline{y}_{i\cdot} = \mu_i + \overline{e}_{i\cdot}$ and $\overline{y}_{\cdot\cdot} = \overline{\mu}_{\cdot} + \overline{e}_{\cdot\cdot}$ into Equation (2A.5), where $\overline{e}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{r_i} e_{ij}$ and $\overline{\mu}_{\cdot} = \frac{1}{N}\sum_{i=1}^{t} r_i\mu_i$. The resulting expression is

$$\begin{aligned} SST &= \sum_{i=1}^{t} r_i(\mu_i + \overline{e}_{i\cdot} - \overline{\mu}_{\cdot} - \overline{e}_{\cdot\cdot})^2 \\ &= \sum_{i=1}^{t} r_i(\mu_i - \overline{\mu}_{\cdot})^2 + \sum_{i=1}^{t} r_i(\overline{e}_{i\cdot} - \overline{e}_{\cdot\cdot})^2 + 2\sum_{i=1}^{t} r_i(\mu_i - \overline{\mu}_{\cdot})(\overline{e}_{i\cdot} - \overline{e}_{\cdot\cdot}) \end{aligned}$$

Expanding the second term of the last expression yields

$$\sum_{i=1}^{t} r_i(\overline{e}_{i\cdot} - \overline{e}_{\cdot\cdot})^2 = \sum_{i=1}^{t} r\overline{e}_{i\cdot}^2 - N\overline{e}_{\cdot\cdot}^2$$

Given

$$E(\overline{e}_{\cdot\cdot}^2) = \frac{1}{N^2}E(\overline{e}_{\cdot\cdot}^2) = \frac{1}{N^2}E(e_{11}^2 + e_{12}^2 + \cdots + e_{rt}^2 + \text{crossproducts}) = \frac{1}{N}\sigma^2$$

and $E(\overline{e}_{i\cdot}^2) = \frac{1}{r_i}\sigma^2$, the expectation of SST is found as

$$\begin{aligned} E(SST) &= \sum_{i=1}^{t} r_i(\mu_i - \overline{\mu}_{\cdot})^2 + \sum_{i=1}^{t} r_i E(\overline{e}_{i\cdot}^2) - NE(\overline{e}_{\cdot\cdot}^2) \\ &= \sum_{i=1}^{t} r_i(\mu_i - \overline{\mu}_{\cdot})^2 + t\sigma^2 - \sigma^2 \\ &= \sum_{i=1}^{t} r_i(\mu_i - \overline{\mu}_{\cdot})^2 + (t-1)\sigma^2 \end{aligned}$$

The expectation of MST is

$$E(MST) = \frac{E(SST)}{(t-1)} = \sigma^2 + \frac{1}{(t-1)}\sum_{i=1}^{t} r_i(\mu_i - \overline{\mu}_{\cdot})^2 \qquad \textbf{(2A.6)}$$

The expected value in Equation (2A.6) can be expressed in terms of the treatment effects by the substitution $\tau_i = (\mu_i - \overline{\mu}_{\cdot})$.

If all $r_i = r$, then let $\theta_t^2 = \sum_{i=1}^{t}(\mu_i - \overline{\mu}_{\cdot})^2/(t-1)$ and

$$E(MST) = \frac{E(SST)}{(t-1)} = \sigma_t^2 + r\theta_t^2 \qquad \textbf{(2A.7)}$$

3 Treatment Comparisons

The analysis of variance and least squares estimates of treatment group means provide the basic information necessary for an in-depth analysis of research hypotheses using methods introduced in this chapter. The methods for an in-depth analysis of the responses to the treatment design include planned contrasts among treatment groups, regression response curves for quantitative treatment factors, selection of the best subset of treatments, comparison of treatments to the control, and all pairwise comparisons among treatment means. All of these methods involve a set of simultaneous decisions to be made by the investigator. This simultaneous statistical inference affects statistical errors of inference. Some of those effects and the control of those errors are discussed in this chapter.

3.1 Treatment Comparisons Answer Research Questions

The relationship between research objectives and treatment design requires us to identify treatments relative to their role in the evaluation of research hypotheses. When an experiment is conducted to answer specific questions, the treatments are selected such that comparisons among the treatments will answer the questions. For example, specific questions can be answered from the meat storage experiment in Chapter 2 about the effect of different storage conditions on the growth of bacteria on meat during storage. The four treatments for the meat storage experiment were (1) commercial wrap, (2) vacuum, (3) mixed gases, and (4) pure CO_2. The summary statistics from the experiment are shown in Display 3.1.

Questions that can be asked about meat storage conditions include

- Is the creation of an artificial atmosphere more effective in reducing bacterial growth than ambient air with commercial wrap?

Display 3.1 Summary Statistics from the Meat Storage Experiment of Example 2.1

	Treatment			
	Commercial	*Vacuum*	*CO, O_2, N*	*CO_2*
$\hat{\mu}_i = \overline{y}_{i.}$	7.48	5.50	7.26	3.36
$t = 4$	$r_i = 3$	$s^2 = MSE = 0.116$ with 8 degrees of freedom		

- Are the gases more effective in reducing bacterial growth than a complete vacuum?
- Is pure CO_2 more effective than a mixture of CO, O_2, and N in reducing bacterial growth?

3.2 Planning Comparisons Among Treatments

Contrasts among treatment means can be constructed to answer specific questions formulated for the experiment. **Contrasts** are special forms of linear functions of observations (discussed in Appendix 3A). A contrast among means is defined as

$$C = \sum_{i=1}^{t} k_i \mu_i = k_1\mu_1 + k_2\mu_2 + \cdots + k_t\mu_t \qquad (3.1)$$

where $\sum_{i=1}^{t} k_i = 0$.

The appropriate contrasts for the three specific questions from the meat storage experiment are

- commercial wrap versus artificial atmospheres:

$$C_1 = \mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$$

- vacuum versus gases:

$$C_2 = \mu_2 - \frac{1}{2}(\mu_3 + \mu_4)$$

- mixed gases versus pure CO_2:

$$C_3 = \mu_3 - \mu_4$$

The first contrast is the difference between the mean of the commercial wrap and the average of the means for the other treatments. The sum of the coefficients $(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$ is zero, as it should be for a proper contrast. The second contrast

is the difference between the mean of the vacuum wrap and the average of the means for the gas treatments with coefficients $(0, 1, -\frac{1}{2}, -\frac{1}{2})$, while the third contrast is a difference between the means of the two gas treatments with coefficients $(0, 0, 1, -1)$.

Estimates of the contrasts, standard errors, confidence interval estimates, and tests of hypotheses about the contrasts are computed from the observed data. Any contrast of the population treatment means, $C = \Sigma k_i\mu_i$, is estimated by the same contrast of the observed treatment means as

$$c = \sum_{i=1}^{t} k_i\overline{y}_{i.} = k_1\overline{y}_{1.} + \cdots + k_t\overline{y}_{t.} \qquad (3.2)$$

where $\overline{y}_{i.}$ is the mean of the ith treatment group.

The estimates for the three contrasts in the meat storage experiment are

$$c_1 = \overline{y}_{1.} - \frac{1}{3}(\overline{y}_{2.} + \overline{y}_{3.} + \overline{y}_{4.}) = 7.48 - \frac{1}{2}(5.50 + 7.26 + 3.36) = 2.11$$

$$c_2 = \overline{y}_{2.} - \frac{1}{2}(\overline{y}_{3.} + \overline{y}_{4.}) = 5.50 - \frac{1}{2}(7.26 + 3.36) = 0.19$$

$$c_3 = \overline{y}_{3.} - \overline{y}_{4.} = 7.26 - 3.36 = 3.90$$

Assessment of Contrast Estimates

Standard Errors of Contrasts

The variance of a contrast estimate $c = \Sigma k_i\overline{y}_{i.}$ is estimated with

$$s_c^2 = s^2\left[\sum_{i=1}^{t}\frac{k_i^2}{r_i}\right] = s^2\left[\frac{k_1^2}{r_1} + \frac{k_2^2}{r_2} + \cdots + \frac{k_t^2}{r_t}\right] \qquad (3.3)$$

where $s^2 = MSE$ from the analysis of variance. If the r_i are equal, the variance estimator simplifies to

$$s_c^2 = \frac{s^2}{r}\left[k_1^2 + k_2^2 + \cdots + k_t^2\right] \qquad (3.4)$$

The estimator for the standard error of a contrast is the square root of the variance,

$$s_c = \sqrt{s_c^2} \qquad (3.5)$$

For the meat storage experiment $MSE = 0.116$ and all $r_i = 3$. The variance and standard error estimates for the first contrast are

$$s_{c_1}^2 = \frac{0.116}{3}\left[1^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2\right] = 0.052$$

and

$$s_{c_1} = \sqrt{0.052} = 0.228$$

The standard error estimates for the second and third contrasts are $s_{c_2} = 0.241$ and $s_{c_3} = 0.278$.

Interval Estimates for Contrasts

The $100(1-\alpha)\%$ confidence interval estimator for a contrast is

$$c \pm t_{\alpha/2,(N-t)}(s_c) \tag{3.6}$$

The correct degrees of freedom for the Student t statistic are the degrees of freedom for the variance used to compute the standard error of the contrast. The variance estimate for the meat storage experiment is $MSE = 0.116$ with $(N-t) = 8$ degrees of freedom. The 95% confidence interval estimate for C_1 using $t_{.025,8} = 2.306$ is

$$2.11 \pm (2.306)(0.228)$$

The 95% interval estimates for the complete set of contrasts are shown in Table 3.1.

Table 3.1 The estimates, standard errors, and 95% confidence interval estimates of three contrasts for the meat storage experiment

		Standard	*95% CI*	
Contrast	*Estimate*	*Error*	*Lower*	*Upper*
C_1	2.11	0.228	1.58	2.64
C_2	0.19	0.241	−0.37	0.75
C_3	3.90	0.278	3.26	4.54

A Sum of Squares Partition for the Contrast

A sum of squares is calculated for the treatment contrast to indicate how much of the variation in the data can be explained by that specific contrast. The sum of squares reduction for estimation of a contrast ($C = \Sigma k_i \mu_i$) can be computed from treatment means as

$$SSC = \frac{\left(\sum_{i=1}^{t} k_i \bar{y}_{i.}\right)^2}{\sum_{i=1}^{t} (k_i^2/r_i)} \tag{3.6}$$

If all r_i are equal

$$SSC = r\frac{\left(\sum_{i=1}^{t} k_i \bar{y}_{i.}\right)^2}{\sum_{i=1}^{t} k_i^2} \tag{3.7}$$

There is 1 degree of freedom associated with the sum of squares for the contrast. The sum of squares partitions for the three contrasts in the meat storage experiment are shown in Table 3.2. The manual computations are illustrated in Display 3.2.

Table 3.2 Analysis of variance with contrasts for log(count/cm^2) of psychrotrophic bacteria from the meat storage experiment

Source	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F*	*Pr > F*
Total	11	33.80			
Treatments	3	32.87	10.96	94.58	.000
Error	8	0.93	0.12		

Contrast	*DF*	*Contrast SS*	*Mean Square*	*F*	*Pr > F*
1 vs others	1	9.99	9.99	86.19	.000
2 vs 3 and 4	1	0.07	0.07	0.62	.453
3 vs 4	1	22.82	22.82	196.94	.000

Hypotheses About Contrasts

The usual null hypothesis states that the contrast has a zero value. For example, the null hypothesis for the second contrast of the meat storage experiment would be

$$H_0\colon C_2 = \mu_2 - \frac{1}{2}(\mu_3 + \mu_4) = 0$$

or, equivalently,

$$H_0\colon C_2 = 2\mu_2 - \mu_3 - \mu_4 = 0$$

If the null hypothesis is true, $C_2 = 0$, the average bacterial growth for the two gas treatments is the same as that for the vacuum treatment.

Display 3.2 Sums of Squares Computations for Contrasts in the Meat Storage Experiment

$$\bar{y}_{1.} = 7.48 \quad \bar{y}_{2.} = 5.50 \quad \bar{y}_{3.} = 7.26 \quad \bar{y}_{4.} = 3.36$$

	k_1	k_2	k_3	k_4
C_1:	3	−1	−1	−1
C_2:	0	2	−1	−1
C_3:	0	0	1	−1

$$SSC_1 = 3[3(7.48) - 5.50 - 7.26 - 3.36]^2/[3^2 + (-1)^2 + (-1)^2 + (-1)^2]$$
$$= 3(6.32)^2/12 = 9.99$$
$$SSC_2 = 3[2(5.50) - 7.26 - 3.36)^2/[2^2 + (-1)^2 + (-1)^2$$
$$= 3(0.38)^2/6 = 0.07$$
$$SSC_3 = 3[7.26 - 3.36]^2/[1^2 + (-1)^2]$$
$$= 3(3.9)^2/2 = 22.82$$

The two forms of the contrasts stated for the null hypotheses are equivalent except the coefficients of the latter differ from the former by a multiple of 2. The choice is purely a matter of personal preference.

The null hypotheses for the three contrasts in the meat storage experiment are

$$H_0: C_1 = 3\mu_1 - \mu_2 - \mu_3 - \mu_4 = 0$$
$$H_0: C_2 = 2\mu_2 - \mu_3 - \mu_4 = 0$$
$$H_0: C_3 = \mu_3 - \mu_4 = 0$$

The first of these hypotheses states that there is no difference between the commercial wrap and the wraps with an artificial atmosphere. The second states that there is no difference between a vacuum and the two gas atmosphere treatments. The third states that there is no difference between pure CO_2 and a mixture of CO, O_2, and N.

Testing Contrasts with the F *Test*

The alternate hypotheses are nonzero differences, or H_a: $C \neq 0$. The null hypothesis, H_0: $C = 0$, is tested with

$$F_0 = \frac{MSC}{MSE} \tag{3.9}$$

where $MSC = SSC$ is the 1 degree of freedom mean square for the contrast. Under the null hypothesis F_0 has the F distribution with 1 and $(N - t)$ degrees of freedom. The null hypothesis is rejected at the α level of significance if $F_0 > F_{\alpha,1,(N-t)}$.

The F_0 test statistic for each of the contrasts is given in the column labeled "F" in Table 3.2. The column labeled "$Pr > F$" shows the probability of exceeding the observed F_0 statistic. The critical value for F_0 at the $\alpha = .05$ level of significance is $F_{.05,1,8} = 5.32$. The null hypothesis is rejected for contrasts C_1 and C_3, $Pr > F = .000$, and not rejected for contrast C_2, $Pr > F = .453$.

Since the contrast C_1 has a positive value estimate, $c_1 = 2.11$, we can conclude that the mean number of bacteria on the meat wrapped in the conventional commercial wrap ($\bar{y}_{1.} = 7.48$) was greater than the average over the meats wrapped in the artificial atmospheres, $\frac{1}{3}(\bar{y}_{2.} + \bar{y}_{3.} + \bar{y}_{4.}) = \frac{1}{3}(5.50 + 7.26 + 3.36) = 5.37$. The contrast C_3 has a positive value estimate, $c_3 = 3.90$, and we can conclude that there are less bacteria on the meat in the CO_2 ($\bar{y}_{4.} = 3.36$) than on the meat in the mixed gas atmosphere ($\bar{y}_{3.} = 7.26$).

Testing Contrasts with the Student t *Test*

Tests of the treatment contrasts can be conducted with the Student t test as well as with the F test. The relationship between the Student t distribution and the F distribution is

$$t^2 = F \tag{3.10}$$

where the Student t has ν degrees of freedom and the F statistic has 1 numerator degree of freedom and ν denominator degrees of freedom. The relationship is always valid when there is 1 degree of freedom in the numerator of the F statistic. The degrees of freedom for the Student t will be the same as that for the denominator of the F statistic.

For any treatment contrast estimate c, with standard error s_c, the ratio

$$t_0 = \frac{c}{s_c} \tag{3.11}$$

has the Student t distribution under the null hypothesis H_0: $C = 0$.

For example, the estimate for the first contrast of the meat storage experiment was $c_1 = 2.11$ with a standard error of $s_{c_1} = 0.228$ (Table 3.1). The ratio

$$t_0 = \frac{2.11}{0.228} = 9.254$$

tests the null hypothesis of no difference between the commercial packaging and the gas atmosphere packagings. The null hypothesis is rejected at the .05 level of significance since $t_0 > t_{.025,8} = 2.306$.

Confidence Intervals or *P*-Values?

The P-value conveniently summarizes with a single value the significance or nonsignificance of hypotheses tests about the contrasts. However, the P-value conveys

no quantitative information about the contrast. A small P-value, indicating significance, may or may not indicate a contrast is meaningfully different from the hypothesized value.

Confidence intervals, conversely, provide meaningful quantitative information about the contrasts as well as their status with respect to the hypotheses about their values. The results from the Amiodarone study in Example 2.2 can be used to illustrate the benefits of confidence intervals.

Recall the estimated increase in the Amiodarone-treated rabbit ears was 1.20°C, a clinically significant value, while those estimates for the vehicle and saline solutions were 0.13°C and 0.00°C, respectively, which were not clinically significant.

If Amiodarone increased the temperature more than the vehicle solution in which it was carried, then Amiodarone would be seen as contributing to tissue inflammation. Likewise, the comparison of the vehicle solution with the saline solution would convey similar information about the contribution of the vehicle solution to tissue inflammation.

The following table shows the two contrasts' estimates along with their standard errors, 95% confidence intervals, and P-values for the Student t test or F test. Case A shows the actual results for the study. Case B shows the situation where the P-values remain the same but the estimates, standard errors, and attendant 95% confidence interval limits have been artificially inflated threefold.

Contrast	*Case*	c	s_c	(L, U)	*P-value*
Amiodarone–Vehicle	A	1.07	0.25	(0.55, 1.59)	.0036
	B	3.21	0.75	(1.65, 4.77)	.0036
Vehicle–Saline	A	0.13	0.25	(−0.39, 0.65)	.6088
	B	0.39	0.75	(−1.17, 1.95)	.6088

The 95% confidence interval estimate for Amiodarone versus Vehicle in Case A indicates Amiodarone significantly elevated the ear temperature over the Vehicle solution with a positive lower limit and an upper limit of 1.59°C, which is clinically significant. Case B shows a threefold increase in the contrast and its standard error with an attendant threefold increase in the confidence interval upper limit to 4.77°C, which is dramatically higher than that for Case A. However, the P-value for both cases is the same at .0036, indicating only that the contrast is different from 0 with no information about the magnitude and direction of the difference.

The 95% confidence interval estimate for Vehicle versus Saline in Case A indicates no difference between the two solutions, with the limits including 0 and both limits less than clinically significant. Case B again shows the artificial threefold increase in the contrast estimate, standard error, and confidence interval limits. However, in this latter case the limits exceed clinical significance. The P-value of .6088 in both cases indicates no significant difference from 0 for the contrast and indicates no more than that. Case A did not approach clinical significance, but it is evident that the limits for the interval estimate of Case B did exceed clinical significance. Thus, with Case B the investigator may want to consider further

action to achieve more precise estimates for the comparison of Vehicle and Saline treatments.

Orthogonal Contrasts Convey Independent Information

A certain class of contrasts, known as **orthogonal contrasts**, has special properties with respect to the sum of squares partitioning in the analysis of variance and with respect to their relationship to one another. *Orthogonality* implies that one contrast conveys no information about the other. Contrast c_2, a comparison between the vacuum packaging and the two gas atmospheres, conveys no information about contrast c_3, a comparison between the two gas atmospheres.

Suppose there are two contrasts c and d, where

$$c = \sum_{i=1}^{t} k_i \overline{y}_{i\cdot} = k_1 \overline{y}_{1\cdot} + k_2 \overline{y}_{2\cdot} + \cdots + k_t \overline{y}_{t\cdot}$$

and

$$d = \sum_{i=1}^{t} d_i \overline{y}_{i\cdot} = d_1 \overline{y}_{1\cdot} + d_2 \overline{y}_{2\cdot} + \cdots + d_t \overline{y}_{t\cdot}$$

The contrasts c and d are orthogonal if

$$\sum_{i=1}^{t} \frac{k_i d_i}{r_i} = \frac{k_1 d_1}{r_1} + \frac{k_2 d_2}{r_2} + \cdots + \frac{k_t d_t}{r_t} = 0 \qquad (3.12)$$

The weighted sum of crossproducts of the coefficients k_i and d_i must sum to 0. If all r_i are equal, then c and d are orthogonal if $\Sigma\, k_i d_i = 0$.

There are $(t - 1)$ mutually orthogonal contrasts among t treatment means; each pair of contrasts will be an orthogonal pair. For example, there are three mutually orthogonal contrasts possible in one set of contrasts among four treatment means.

The coefficients for two of the contrasts of the meat storage experiment, c_2 and c_3, with equal treatment replications are shown in Display 3.3 along with the crossproducts of the coefficients. The two contrasts are orthogonal because their crossproducts sum to 0. Analogous calculations show that the contrast c_1, $(3, -1, -1, -1)$, is also orthogonal to c_2 and c_3.

Display 3.3 Orthogonal Contrasts for the Meat Storage Experiment

	Coefficients			
Contrast	k_1	k_2	k_3	k_4
c_2	0	2	−1	−1
c_3	0	0	1	−1

Sum of crossproducts $= (0)(0) + (2)(0) + (-1)(1) + (-1)(-1) = 0$

Another special feature of orthogonal contrasts is that the sums of squares for the $(t-1)$ orthogonal contrasts sum to the treatment sum of squares in the analysis of variance. The sums of squares for the three orthogonal contrasts of the example (c_1, c_2, and c_3) sum to the treatment sum of squares within rounding errors, $SST = 32.87$ as seen in Table 3.2.

Contrasts Among Treatments with Unequal Replication

The Amiodarone experiment in Example 2.2 had unequal replication of the treatment groups. The summary statistics for the experiment are shown in Display 3.4.

Display 3.4 Summary Statistics from the Amiodarone Experiment of Example 2.2

	Treatment		
	Amiodarone	*Vehicle*	*Saline*
$\widehat{\mu}_i = \overline{y}_{i.}$	1.20	0.13	0.00
r_i	9	6	8

$t = 3 \quad N = 23 \quad s^2 = MSE = 0.2177$ with 20 degrees of freedom

Two comparisons of interest may be (1) the contrast of the Amiodarone treatment mean with the vehicle and saline means and (2) the contrast between the vehicle and saline means. The two contrasts, respectively, are estimated by

$$c_1 = 2\overline{y}_{1.} - \overline{y}_{2.} - \overline{y}_{3.}$$

and

$$c_2 = \overline{y}_{2.} - \overline{y}_{3.}$$

The contrast coefficients are $(2, -1, -1)$ and $(0, 1, -1)$ with replication numbers $(9, 6, 8)$. Evaluating the contrasts using Equation (3.12),

$$\frac{2(0)}{9} + \frac{-1(1)}{6} + \frac{-1(-1)}{8} \neq 0$$

indicates that the two contrasts are not orthogonal.

For the contrasts to be orthogonal the values of the coefficients would have to be altered. For example, two contrasts for the Amiodarone experiment that satisfy the orthogonality criterion in Equation (3.12) are

$$c_1 = 7\overline{y}_{1.} - 3\overline{y}_{2.} - 4\overline{y}_{3.}$$

and

$$c_2 = \overline{y}_{2.} - \overline{y}_{3.}$$

The coefficients $(7, -3, -4)$ and $(0, 1, -1)$ satisfy the condition of a contrast with $\Sigma k_i = 0$. Evaluation by Equation (3.12) is

$$\frac{7(0)}{9} + \frac{-3(1)}{6} + \frac{-4(-1)}{8} = 0$$

The two contrasts would now be orthogonal and convey no information about one another.

However, consider the first contrast, $c_1 = 7\overline{y}_{1.} - 3\overline{y}_{2.} - 4\overline{y}_{3.}$. The coefficients imply a weighted comparison among the population means. Unless the treatment populations occur in the proportions 7:3:4 the comparison does not make sense. There is unequal information in each of the treatment groups within the experiment; however, it is not necessarily so that the treatment populations would be unequally represented in nature. In the case of the Amiodarone experiment the contrast with equal weights for the treatment means is a more sensible comparison.

Comments on Orthogonality

To present a nicely ordered analysis of variance table, the choice of contrasts for a study must not be dictated by their orthogonality. Rather, the contrasts should be constructed to answer specific research questions. For example, it may be of interest in the meat storage experiment to contrast the control treatment separately with each of the other treatments, resulting in a non-orthogonal set of contrasts, $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_1 - \mu_4$. Orthogonal contrasts purposely were chosen for the meat storage experiment to simplify the presentation. However, they did in the meantime answer some specific meaningful questions, which is after all the purpose of a research study. The research hypotheses and the treatment design should dictate the construction of the contrasts.

3.3 Response Curves for Quantitative Treatment Factors

Many studies are conducted to determine the quantitative trend relationship between two variables. In experimental studies, one of the variables is usually under the *control* of the investigator while the other is the *observed response* variable. The factorial treatment designs discussed in Section 1.4 are an example of structured treatment designs in which a quantitative factor may have several graded levels.

For example, an experiment can be designed to measure the growth response of animals to increasing amounts of nutrient in the diet. The treatment design consists of one quantitative factor—the amount of nutrient in the diet with three levels, say, 0, 500, and 1000 parts per million (ppm). The objective of the study will be to characterize animal growth as a function of the amount of nutrient in the diet.

Polynomial Response Functions for Response Curves

The polynomial model often used to describe trend relationships between a measured response y and quantitative levels of a factor x is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + e \qquad \textbf{(3.13)}$$

Several examples of response curves are shown in Figure 3.1. The linear, quadratic, and cubic polynomial response curves provide good approximations to many relationships common to the biological and physical sciences.

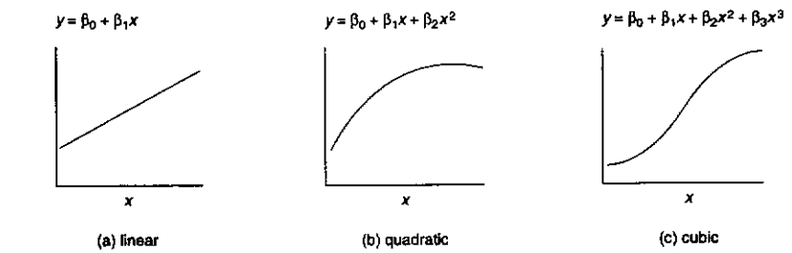

Figure 3.1 Examples of polynomial response curves: (a) linear, (b) quadratic, and (c) cubic

Example 3.1 Grain Production and Plant Density

The growth and development of plants, as with any living organism, is dependent on the availability of sufficient nutrition. The objective for most cultivated grain crops is to maximize seed or grain production under a given cultivation and soil fertility regimen. If the grain producer plants too few seeds per unit of area, maximum grain production will not be realized simply because each plant has genetically determined maximum potential. Thus, increased grain production requires more plants per unit area. On the other hand, for an entirely different reason, maximum production will not be realized if an excessive number of plants are grown per unit of area. Available nutrition per plant becomes limiting, and the plant growth and development are reduced with concomitant reduction in grain production.

The objective for a crop scientist under these circumstances is to characterize the relationship between plants per unit of area and grain production under a given cultivation and fertility regimen. An experiment was conducted to estimate a polynomial response curve to characterize the relationship.

The treatment design consisted of five plant densities (10, 20, 30, 40, and 50). Each of the five treatments was assigned randomly to three field plots in a completely randomized experiment design. The resulting grain yields are shown in Table 3.3.

Table 3.3 Amount of grain produced per plot for five plant densities

	Plant Density (x)				
	10	*20*	*30*	*40*	*50*
	12.2	16.0	18.6	17.6	18.0
	11.4	15.5	20.2	19.3	16.4
	12.4	16.5	18.2	17.1	16.6
Means ($\overline{y}_{i.}$)	12.0	16.0	19.0	18.0	17.0

A graph of the observations in Figure 3.2 suggests a quadratic relationship between grain yield and plant density. The mean responses of grain yields ($\overline{y}_{i.}$) to plant density (x) shown in Table 3.3 also suggest a quadratic response polynomial as an approximate model to describe the biological relationship between grain yield and plant density for this study. The objective is to determine the lowest possible order polynomial equation that adequately describes the relationship. With five values of x it is possible to fit a fourth-degree equation in x; however, a simpler and less complex second-degree equation may do just as well.

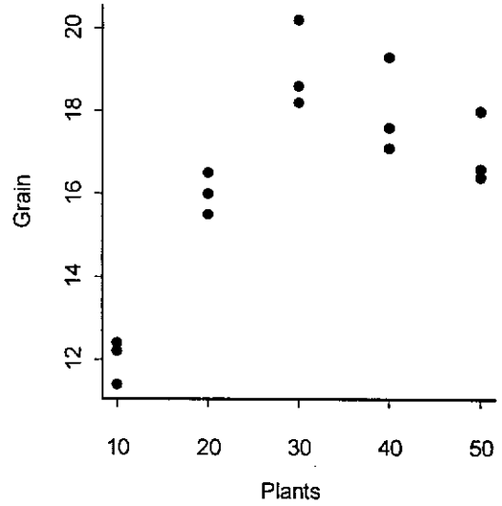

Figure 3.2 Grain yield versus plant density

Simplifying with Orthogonal Polynomials

The polynomial model can be fit to the observed data with many of the available computer regression programs. The trend analysis can be simplified by examining orthogonal contrasts among the treatment factor levels that measure the linear, quadratic, and higher level polynomial effects. These contrasts, known as **orthogonal**

polynomials, enable us to evaluate the importance of each polynomial component with a specific contrast.

In Example 3.1, there are $t = 5$ levels of the plant density factor, and $(t - 1) = 4$ orthogonal contrasts can be estimated. After transforming the polynomials in x into orthogonal polynomials, the complete orthogonal polynomial equation model for the relationship between plant density and grain yield is

$$y_{ij} = \mu + \alpha_1 P_{1i} + \alpha_2 P_{2i} + \alpha_3 P_{3i} + \alpha_4 P_{4i} + e_{ij} \qquad \textbf{(3.14)}$$

where μ is the grand mean and P_{ci} is the cth-order orthogonal polynomial for the ith level of the treatment factor.

The transformations for the powers of x into orthogonal polynomials (P_{ci}) up to the third degree are shown in Display 3.5. The tabled values of the orthogonal polynomials for $t = 3$ to 10 are given in Appendix Table XI. The values from Display 3.5 or Appendix Table XI are valid for any distance between the values of x as long as the spacing is equal between all values of x (or factor levels) and replication numbers are the same for all treatments. The constant λ_i at the beginning of each transformation makes each of the P_i values an integer value.

Display 3.5 Transformation of the Powers of x into Orthogonal Polynomials

Mean: $P_0 = 1$

Linear: $P_1 = \lambda_1 \left[\dfrac{x - \overline{x}}{d} \right]$

Quadratic: $P_2 = \lambda_2 \left[\left(\dfrac{x - \overline{x}}{d} \right)^2 - \left(\dfrac{t^2 - 1}{12} \right) \right]$

Cubic: $P_3 = \lambda_3 \left[\left(\dfrac{x - \overline{x}}{d} \right)^3 - \left(\dfrac{x - \overline{x}}{d} \right) \left(\dfrac{3t^2 - 7}{20} \right) \right]$

t = number of levels of the factor — x = value of the factor level
$\overline{x}$ = mean of the factor levels — d = distance between factor levels

Note the term $(x - \overline{x})/d$ occurs consistently in all of the orthogonal polynomial transformations in Display 3.5. The transformation $(x - \overline{x})$ centers the x values around 0, while dividing the result by the increment between the values of x scales the values to change by one unit between levels. For example, in Example 3.1 with $\overline{x} = 30$ and $d = 10$ the resulting transformation is

x:	10	20	30	40	50
$(x - 30)$:	−20	−10	0	10	20
$(x - 30)/10$:	−2	−1	0	1	2

Some of the standard computer program packages for statistical analysis are capable of producing the orthogonal polynomials with unequal or equal spacing between the x values. Grandage (1958) gave a manual method for deriving the orthogonal polynomials with unequal spacings.

The orthogonal polynomial transformations for the plant densities are shown in Table 3.4 with $d = 10$, $t = 5$, $\overline{x} = 30$, and $x = 10$, 20, 30, 40, or 50. Each set of coefficients P_1 through P_4 forms a contrast among the treatments since the sum of the coefficients in each of these columns is equal to 0. The contrasts are mutually orthogonal.

Table 3.4 Computations for orthogonal polynomial contrasts and sums of squares

		Orthogonal Polynomial Coefficients (P_{ci})				
Density (x)	$\overline{y}_{i.}$	*Mean*	*Linear*	*Quadratic*	*Cubic*	*Quartic*
10	12	1	−2	2	−1	1
20	16	1	−1	−1	2	−4
30	19	1	0	−2	0	6
40	18	1	1	−1	−2	−4
50	17	1	2	2	1	1
λ_c		–	1	1	5/6	35/12
Sum $= \Sigma P_{ci}\overline{y}_{i.}$		82	12	−14	1	7
Divisor $= \Sigma P_{ci}^2$		5	10	14	10	70
$SSP_c = r(\Sigma P_{ci}\overline{y}_{i.})^2/\Sigma P_{ci}^2$		–	43.2	42.0	0.3	2.1
$\widehat{\alpha}_c = \Sigma P_{ci}\overline{y}_{i.}/\Sigma P_{ci}^2$		16.4	1.2	−1.0	0.1	0.1

The treatment sum of squares can be partitioned into an additive set of 1 degree of freedom sums of squares, one sum of squares for each of the $(t - 1)$ orthogonal polynomial contrasts. Consequently, it is possible to test sequentially the significance of the linear, quadratic, cubic, and so forth, terms in the model to determine the best-fitting polynomial equation.

The estimates of the α_c coefficients for the orthogonal polynomial equation in Equation (3.14) and the sum of squares for each of the orthogonal polynomial contrasts are shown in Table 3.4. The estimated orthogonal polynomial equation is found by substituting the estimates of μ, α_1, α_2, α_3, and α_4 from Table 3.4 into Equation (3.14). The estimated equation is

$$\widehat{y}_i = 16.4 + 1.2P_{1i} - 1.0P_{2i} + 0.1P_{3i} + 0.1P_{4i} \qquad \textbf{(3.15)}$$

The analysis of variance for this experiment is shown in the top analysis of Table 3.5. The ratio $F_0 = MST/MSE$ tests the global null hypothesis H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$. At the .05 level of significance the critical region is $F_0 > F_{.05,4,10} = 3.48$. The null hypothesis is rejected since $F_0 = 29.28$ exceeds the critical value.

The 1 degree of freedom sum of squares partitions for each of the orthogonal polynomial contrasts are summarized in the analysis of variance at the bottom of

Table 3.5. Notice the sums of squares for the four contrasts sum to the sum of squares for Density, 87.60 with 4 degrees of freedom.

Table 3.5 Analysis of variance for the orthogonal polynomial model relationship between plant density and grain yield

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F*	*Pr > F*
Density	4	87.60	21.90	29.28	.000
Error	10	7.48	0.75		
Contrast	*DF*	*Contrast SS*	*Mean Square*	*F*	*Pr > F*
Linear	1	43.20	43.20	57.75	.000
Quadratic	1	42.00	42.00	56.15	.000
Cubic	1	.30	.30	.40	.541
Quartic	1	2.10	2.10	2.81	.125

We are interested in the contribution of the separate polynomial terms of the model. One strategy to determine the best polynomial equation is to test the significance of the terms in the sequence: linear, quadratic, cubic, and so forth. Beginning with the simplest polynomial, a more complex polynomial is constructed as the data requires for adequate description.

The sequence of hypotheses is H_0: $\alpha_1 = 0$, H_0: $\alpha_2 = 0$, H_0: $\alpha_3 = 0$, and so forth. These hypotheses about the orthogonal polynomial contrasts are each tested with the F test for the respective contrasts. The ratios $F_0 = MSC/MSE$ are given in Table 3.5 for each of the polynomial contrasts estimated for the plant density study. For each sum of squares partition the null hypothesis is H_0: $\alpha_C = 0$ with critical region $F_0 > F_{.05,1,10} = 4.96$.

The null hypothesis is rejected for the linear and quadratic terms of the model ($Pr > F = .000$) but is not rejected for the cubic ($Pr > F = .541$) and quartic ($Pr > F = .125$) terms. The quadratic model is sufficient for a description of the relationship between grain yield and plant density on the basis of the statistical tests.

Computing the Response Curve

The estimated quadratic response curve without the cubic and quartic terms, $\widehat{\alpha}_3 P_3$ and $\widehat{\alpha}_4 P_4$, from Equation (3.15) is

$$\widehat{y}_i = \overline{y}_{..} + \widehat{\alpha}_1 P_{1i} + \widehat{\alpha}_2 P_{2i} = 16.4 + 1.2 P_{1i} - 1.0 P_{2i} \qquad \textbf{(3.16)}$$

The estimated value for a plant density of $x = 10$ is determined by substituting $P_1 = -2$ and $P_2 = 2$ into Equation (3.16) as $\widehat{y} = 16.4 + 1.2(-2) - 1.0(2) = 12.0$. The observed grain yields and those estimated from Equation (3.16) are shown for all plant densities in Table 3.6.

Table 3.6 Observed grain yields and those estimated from the quadratic orthogonal polynomial equation

Density	*Coefficients*		*Estimated*	*Observed*
x	P_1	P_2	$\widehat{y}_i$	$\overline{y}_{i.}$
10	−2	2	12.0	12
20	−1	−1	16.2	16
30	0	−2	18.4	19
40	1	−1	18.6	18
50	2	2	16.8	17

The polynomial relationship expressed as a function of y and x in actual units of the observed variables is more informative than when expressed in units of the orthogonal polynomial. A direct transformation to an equation in x requires the information in Display 3.5 and Table 3.4. The necessary quantities are $\lambda_1 = 1$, $\lambda_2 = 1$, $d = 10$, $\overline{x} = 30$, and $t = 5$. Then substituting for the P_i

$$\begin{aligned}\widehat{y} &= 16.4 + 1.2P_1 - 1.0P_2 \\ &= 16.4 + 1.2(1)\left[\frac{x-30}{10}\right] - 1.0(1)\left[\left(\frac{x-30}{10}\right)^2 - \left(\frac{5^2-1}{12}\right)\right]\end{aligned}$$

and simplifying to

$$\widehat{y} = 5.8 + 0.72x - 0.01x^2 \qquad \textbf{(3.17)}$$

The estimated line from Equation (3.17) may be plotted as shown in Figure 3.3, along with the observed data points (squares) and treatment means $\overline{y}_{i.}$ (filled circles).

Standard Errors and Confidence Intervals for the Response Curve

The estimated response curve is composed of estimates for several parameters. The quadratic equation chosen as the best-fitting equation for the relationship between grain production and plant density has three estimates: $\overline{y}_{..} = 16.4$, $\widehat{\alpha}_1 = 1.2$, and $\widehat{\alpha}_2 = -1.0$.

The estimator for the variance of $\widehat{y} = \overline{y}_{..} + \widehat{\alpha}_1 P_1 + \widehat{\alpha}_2 P_2$ is

$$s^2_{\widehat{y}} = s^2_{\overline{y}_{..}} + P_1^2 s^2_{\widehat{\alpha}_1} + P_2^2 s^2_{\widehat{\alpha}_2} \qquad \textbf{(3.18)}$$

The variance estimator for a polynomial contrast s_c^2 is

$$s_c^2 = \frac{s^2}{(r\Sigma P_{ci}^2)} \qquad \textbf{(3.19)}$$

and the standard error estimator for a contrast is the square root of the variance

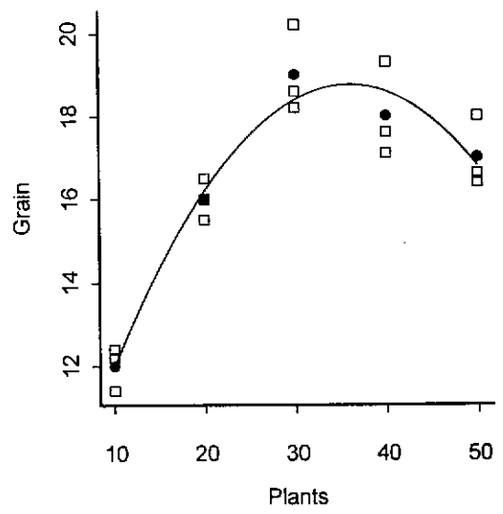

Figure 3.3 Estimated response curve, $\widehat{y} = 5.8 + 0.72x - 0.01x^2$, for grain yield as a function of plant density

$$s_c = \sqrt{s_c^2} \tag{3.20}$$

The variance estimates for the coefficients of the estimated quadratic equation are computed using $s^2 = MSE = 0.75$ from the analysis of variance in Table 3.5 and the divisors from Table 3.4. The variances are

$$s_{\overline{y}_{..}}^2 = \frac{0.75}{(3)(5)} = 0.050$$

$$s_{\widehat{\alpha}_1}^2 = \frac{0.75}{(3)(10)} = 0.025$$

and

$$s_{\widehat{\alpha}_2}^2 = \frac{0.75}{(3)(14)} = 0.018$$

The variance of an estimated value is

$$\begin{aligned} s_{\widehat{y}}^2 &= s_{\overline{y}_{..}}^2 + s_{\widehat{\alpha}_1}^2 P_1^2 + s_{\widehat{\alpha}_2}^2 P_2^2 \\ &= .05 + 0.025P_1^2 + 0.018P_2^2 \end{aligned} \tag{3.21}$$

Since the values of P_1 and P_2 differ for each value of plant density, the variance of the estimated value differs for each value of plant density. The variance of

the estimated yield for plant density $x = 10$, using $P_1 = -2$ and $P_2 = 2$ in Equation (3.21), is

$$s_{\widehat{y}}^2 = 0.05 + 0.025(-2)^2 + 0.018(2)^2 = 0.222$$

and the standard error is

$$s_{\widehat{y}} = \sqrt{0.222} = 0.471$$

The $100(1 - \alpha)\%$ confidence interval for the estimated value is computed from

$$\widehat{y} \pm t_{\alpha/2,(N-t)}(s_{\widehat{y}})$$

The standard errors and 95% confidence intervals for the estimated values of grain yield for all five plant densities using $t_{.025,10} = 2.228$ are shown in Table 3.7.

Table 3.7 Observed and estimated grain yields and standard errors for estimated grain yields for each plant density

	Seed Yield		*Standard*	*95% Confidence*
Density	*Observed*	*Estimated*	*Error*	*Interval*
10	12	12.0	0.471	(10.95, 13.05)
20	16	16.2	0.305	(15.52, 16.88)
30	19	18.4	0.349	(17.62, 19.18)
40	18	18.6	0.305	(17.92, 19.28)
50	17	16.8	0.471	(15.75, 17.85)

The estimated response curve (Figure 3.3) has the advantage of portraying the relationship between y and x throughout the entire range of x values used in the experiment. With this example, it is possible to describe or estimate the grain yield for any plant density between 10 and 50 plants. The description of the relationship is not constrained, in a discussion of the results, to the five plant densities used in the study.

3.4 Multiple Comparisons Affect Error Rates

The group of contrasts exhibited in Section 3.2 are considered **multiple comparisons** because more than one comparison was made among the treatment means. Any necessary number of these comparisons can be constructed in the analysis to help answer the research questions.

However, multiple comparisons among treatment means can lead the investigator, sometimes unwittingly, into a statistical minefield. The abundant number of available procedures increases the difficulty on the part of an investigator of choosing the appropriate method for a particular situation.

Error Rates for Multiple Comparisons

The difficulties with multiple comparisons reside mainly in an understanding of the error rates associated with testing multiple hypotheses. Hypothesis tests have risks associated with decisions to reject or not reject the hypothesis. For any contrast among means the risk associated with declaring the contrast to be real when it is not is the risk of a Type I error. The risk of declaring the contrast among population means to be equal to zero when it is not is the risk of a Type II error.

The risks for Type I and Type II errors are inversely related to one another. The level of significance chosen for the hypothesis test determines the risk of Type I error. Sample size, variance, and size of the contrast for the true population means determine the Type II error rate for a given Type I error rate.

A simple test for the difference between two treatment means is the Student t test, with the statistic calculated as

$$t_0 = \frac{\overline{y}_i - \overline{y}_j}{\sqrt{s^2\left[\frac{1}{r_i} + \frac{1}{r_j}\right]}} \tag{3.22}$$

The significance level or probability of Type I error for a single test is a *comparisonwise* error rate, α_C. It is the risk we are willing to take on a single comparison.

There are $p(p-1)/2$ pairwise comparisons among p treatment means. For example, four treatments (A, B, C, D) have $4(3)/2 = 6$ possible pairs: (A, B), (A, C), (A, D), (B, C), (B, D) and (C, D). If the six pairs of means are tested with the t_0 statistic in Equation (3.22), there is the possibility of committing 0, 1, 2, 3, 4, 5, or 6 Type I errors if the six population means are all equal.

With the possibility of up to six Type I errors for six tests we can use another form of Type I error based on the accumulated risks associated with the family of tests under consideration. The family is the set of pairwise comparisons for the example in the previous paragraph. The accumulated risks associated with a family of comparisons is often called the *experimentwise* Type I error rate, α_E. It is the risk of making at least one Type I error among the family of comparisons in the experiment.

Evaluating the Maximum Error Rate

The experimentwise Type I error rate can be evaluated for a family of independent tests. However, all pairwise tests using Equation (3.22) are not independent since the s^2 in the denominator of each of the t_0 statistics is the same, and the numerator of each test contains the same means as several of the other t_0 statistics.

Although the set of tests in the family just described are not independent, the upper limit for the value of the experimentwise Type I error rate can be evaluated by assuming independent tests. Suppose the null hypotheses are true for each of n independent tests. The probability of a Type I error for any single test is α_C (the comparisonwise rate) with $(1 - \alpha_C)$ as the probability of a correct decision. The

probability of committing x Type I errors is given by the binomial distribution as

$$P(x) = \frac{n!}{x!(n-x)!}\alpha_C^x(1-\alpha_C)^{n-x} \tag{3.23}$$

for $x = 0, 1, 2, \ldots, n$ Type I errors. The probability of having no Type I errors is

$$P(x = 0) = (1 - \alpha_C)^n$$

The probability of committing *at least one* Type I error ($x = 1, 2, 3, \ldots, n$) is $P(x \geq 1) = 1 - P(x = 0)$, or

$$\alpha_E = 1 - (1 - \alpha_C)^n \tag{3.24}$$

The probability α_E is the risk of making at least one Type I error among the n independent comparisons. It is the upper limit of the experimentwise Type I error rate for n tests among a set of treatment means.

The relationship

$$\alpha_C = 1 - (1 - \alpha_E)^{1/n} \tag{3.25}$$

expresses the comparisonwise Type I error rate as a function of the experimentwise Type I error rate.

The relationship between the two Type I error rates for a few selected values of n is shown in Table 3.8. If each of the tests is conducted with a comparisonwise error rate of $\alpha_C = .05$ the risk of at least one Type I error escalates as the number of tests increases. When $n = 1$ test is conducted both Type I errors are identical as they should be since only one Type I error can be committed. When $n = 5$ the risk probability of at least one Type I error among the five decisions has risen to $\alpha_E = .226$, and with $n = 10$ the risk rises to a probability of .401 to commit at least one Type I error.

The last column of Table 3.8 gives an indication of the comparisonwise Type I error rate required to maintain an experimentwise Type I error rate $\alpha_E = .05$. For example, when five independent tests are conducted and we want to keep the risk of committing at least one Type I error as low as 1 in 20 chances, the comparisonwise error rate for each of the $n = 5$ tests must be $\alpha_C = .01$ and for $n = 10$ tests it must be $\alpha_C = .005$.

Table 3.8 The relationship between α_C and α_E for $n = 1, 2, 3, 4, 5$, or 10 independent tests

n	α_E *when* $\alpha_C = .05$	α_C *when* $\alpha_E = .05$
1	.050	.050
2	.098	.025
3	.143	.017
4	.185	.013
5	.226	.010
10	.401	.005

Which Experimentwise Error Rate?

Some confusion exists regarding error rates for simultaneous inference since the error rate can be defined relative to the configuration of the population means, $\mu_1, \mu_2, \ldots, \mu_t$. The *experimentwise* error rate has often been defined under the configuration $\mu_1 = \mu_2 = \cdots = \mu_t$, and under this configuration the familywise error rate is considered controlled in the *weak sense* by Hochberg and Tamhane (1987). Equality of means is unlikely to be true under most circumstances. Therefore, when inequalities do occur the weak sense experimentwise error rate offers poor protection against incorrect decisions. If no constraints are put on the relationships among the μ_i, then the familywise error rate is controlled in the *strong sense* (Hochberg & Tamhane, 1987). The latter *strong sense* definition can be interpreted as the probability of at least one incorrect decision over all parameter configurations. Hsu (1996) also presents a detailed discussion of error rates for simultaneous inference.

3.5 Simultaneous Statistical Inference

The discussion in the previous section considered multiple simultaneous inferences and the probabilities of incorrect decisions regarding those inferences. When hypotheses were tested about three contrasts in the meat storage experiment of Section 3.1, they each were tested with a comparisonwise Type I error rate of $\alpha_C = .05$. If those statements are to hold simultaneously with a Type I error rate of $\alpha_E = .05$, the three tests of hypotheses require a family error rate. According to Table 3.8, the comparisonwise error rate would have to be reduced for the three statements to hold simultaneously with a family error rate of $\alpha_E = .05$. If the contrasts were independent, then $\alpha_C = .017$ would provide an appropriate family error rate under a null hypothesis of equality among the means in the contrasts.

An investigator wants to make informed decisions from the observed data with a valid statistical method. The choice of method depends on the *type* of inference desired and the *strength* of inference desired.

The **types** of contrasts most often considered by investigators are

- planned contrasts
- orthogonal polynomial contrasts
- multiple comparisons with the best treatment
- multiple comparisons with the control treatment
- all pairwise comparisons

The first two methods were discussed in Sections 3.2 and 3.3. The other three methods are considered in later sections in this chapter.

Strength of inference refers to how much can be said about a comparison. The strongest inferences include statements about the direction and magnitude of the difference. *Simultaneous confidence interval* methods provide the strongest form of inference. The investigator is able to assert, with a given level of confidence, in which direction the effects of one treatment differ from another and also how far removed those effects are from one another.

Methods providing *confident directions* inference follow simultaneous confidence intervals in strength (Hsu, 1996). These methods, for a given contrast C_i, assert inequalities ($C_i > 0$ or $C_i < 0$) for each contrast with given levels of confidence that all statements are correct but say nothing about magnitude.

Confident inequalities declare the inequality for each contrast as $C_i \neq 0$ with a given confidence level that all declarations are correct (Hsu, 1996). These statements make no assertion about the direction or the magnitude of the inequality.

The weakest methods are *individual comparison* methods that do not consider the simultaneous nature of the inference. The weakness of these methods was demonstrated in the previous section where in Table 3.8 the probability of error for simultaneous statements, α_E, could increase considerably as the number of tests increased.

Multiple Comparison Methods, Probability Tables, and Computer Programs

Well-defined methods with explicit probability tables exist for some of the multiple comparison procedures discussed later in this chapter, given the study can be modeled as a completely randomized design with equal replication numbers for treatments.

Either approximations to the probability tables or computer programs must be employed if these methods are used with unequal replications or more complex design structures.

Fortunately, a number of statistical program suites include exact computations for these methods. However, the documentation for the programs should be checked to be certain valid methods are being used.

Approximations to the probability tables take advantage of several inequalities from probability theory. If the special computer routines are not available these approximations will provide a conservative alternative. The approximation that will be suggested for methods later in this chapter is based on the **Bonferroni inequality** because probability tables based on the inequality are readily available. This approximation can also be used for the less structured methods, such as small sets of planned contrasts and the orthogonal polynomial contrasts, to provide a measure of error protection for simultaneous inference.

Bonferroni *t* Statistics for Simultaneous Inference

The *Bonferroni inequality* provides a means to obtain an easy approximation to multiple comparison error rates. The inequality, translated to our context, shows that the family error rate is less than or equal to the sum of the individual comparison error rates. When n comparisons are made at the same comparison error rate, α_C, the Bonferroni inequality gives the relationship

$$\alpha_E \leq n\alpha_C$$

Equality of the relationship holds when the tests are independent. The comparisonwise error rate for the test statistic is determined by dividing the maximum desired family error rate by the number of simultaneous tests, $\alpha_C = \alpha_E/n$. For example, with three comparisons for the meat storage experiment a family error rate of $\alpha_E = .05$ requires a comparisonwise error rate of $\alpha_C = .05/3 = .017$.

Tabled values of the Student t statistic, referred to as the Bonferroni t, for selected values of α_E are given in Appendix Table V as $t_{\alpha_E/2,k,\nu}$ for two-sided tests where k is the number of comparisons and ν is the degrees of freedom. Values can also be obtained from any computer program that can compute probabilities for the Student t distribution for upper-tail probability $\alpha_E/2k$. For example, the Bonferroni t for $k = 3$ two-sided comparisons with $\nu = 8$ degrees of freedom and $\alpha_E = .05$ is $t_{.025,3,8} = 3.02$. Equivalently, the value can be found by determining the value of the Student t for $\nu = 8$ degrees of freedom exceeded with probability $\alpha_E/2k = .05/2(3) = .00833$.

Simultaneous Confidence Interval (SCI) Estimates

Confidence interval estimates for contrasts are interval estimates that hold simultaneously with confidence level $100(1 - \alpha_E)\%$. The simultaneous confidence intervals use the Bonferroni t statistic in place of the usual Student t statistic. The two-sided $100(1 - \alpha_E)\%$ confidence interval is

$$c \pm t_{\alpha_E/2,k,\nu}(s_c) \qquad \textbf{(3.26)}$$

The three contrasts for the meat storage experiment require $t_{.025,3,8} = 3.02$ for a set of intervals that hold simultaneously with at least 95% confidence. The resulting intervals are shown in Display 3.6.

Display 3.6 Simultaneous 95% Confidence Intervals for Three Contrasts in the Meat Storage Experiment

Contrast	*Estimate*	*Standard Error*	*95% SCI (L, U)*
$\mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$	2.11	0.228	(1.42, 2.80)
$\mu_2 - \frac{1}{2}(\mu_3 + \mu_4)$	0.19	0.241	(−0.54, 0.92)
$\mu_3 - \mu_4$	3.90	0.278	(3.06, 4.74)

The SCI intervals are wider than those computed as individual 95% confidence intervals in Table 3.1. There is a trade-off between the strength of the confidence statement and the interval widths. The joint confidence level is less for the shorter intervals in Table 3.1 and greater for the wider SCI intervals shown in Display 3.6.

The investigator can make joint statements with a Type I experimentwise error rate of .05 that (1) the average bacterial growth in artificial atmospheres is less than that on the meat in the commercial wrap since the lower limit of the interval is greater than 0; (2) there is no difference between the vacuum wrap and the average of the gases with respect to the bacterial growth since the interval includes 0; and (3) pure CO_2 results in fewer bacteria than the mixed gas since the lower limit of the interval is greater than 0.

The three contrasts for the meat storage experiment from Table 3.1, their estimates, standard errors, and calculated t_0 statistics are shown in Display 3.7. The calculated t_0 statistics for the first and third contrasts exceed the critical value $t_{.025,3,8} = 3.02$. The t statistics in Display 3.7 exemplify the use of *confident inequalities* inference if only their significance is evaluated. The magnitudes and signs of the contrasts and the t statistics can be used to deduce the direction and magnitude of the comparisons, although in a less direct manner than with the confidence intervals.

Display 3.7 Bonferroni t Tests for Three Contrasts in the Meat Storage Experiment

Contrast	*Estimate*	*Standard Error*	t_0
C_1 (Commercial vs. artifical)	2.11	0.228	9.25
C_2 (Vacuum vs. gases)	0.19	0.241	0.79
C_3 (CO_2 vs. mixed gas)	3.90	0.278	14.03

Scheffé's Test for Simultaneous Inference

Bonferroni t statistics can be used safely for a small number of pre-planned contrasts with preservation of the proposed experimentwise error. A method for testing *all possible* contrasts or constructing confidence intervals for *all possible* contrasts was proposed by Scheffé (1953). The method provides the prescribed experimentwise error protection for any number of contrasts. Consequently, the method is quite conservative and is generally used for unplanned contrasts or contrasts suggested by the data. The Scheffé test is shown in Display 3.8.

Simultaneous $100(1 - \alpha_E)\%$ confidence intervals for all possible contrasts are computed with the Scheffé statistic as

$$c \pm S(\alpha_E) \qquad \textbf{(3.27)}$$

and there is a $(1 - \alpha_E)$ probability that all intervals simultaneously include the true values of the respective contrasts.

Display 3.8 The Scheffé Test

Consider any contrast, $c = \sum_{i=1}^{t} k_i \overline{y}_{i.}$, among t treatment means with standard error

$$s_c = \sqrt{s^2 \left[\sum_{i=1}^{t} \frac{k_i^2}{r_i} \right]}$$

The null hypothesis for the contrast, H_0: $C = 0$, is rejected if

$$| c | > S(\alpha_E) \qquad \textbf{(3.28)}$$

$S(\alpha_E)$ is the Scheffé statistic

$$S(\alpha_E) = s_c \sqrt{(t-1) F_{\alpha_E,(t-1),\nu}} \qquad \textbf{(3.29)}$$

where $F_{\alpha_E,(t-1),\nu}$ is the F statistic with $(t-1)$ and ν degrees of freedom exceeded with probability α_E. Also, ν is the number of degrees of freedom for experimental error variance s^2, used to estimate the standard error of the contrast, s_c.

3.6 Multiple Comparisons with the Best Treatment

In some studies, the treatments are not highly structured in their relationship to one another and structured contrasts are difficult to identify. On the other hand, treatments are related to one another because they all are under investigation for their effect on the measured response variables, and they address some specific problem of interest to the investigator. Some possible examples include testing toxins, sources of protein for diets, or mixtures of compounds for an alloy.

Under these circumstances the investigator may want to "pick the winners." The objective is to select the set of treatments or single treatment (if possible) that provides the most desirable result.

The **multiple comparisons with the best (MCB)** procedure from Hsu (1984) enables the investigator to select treatments into a subset such that the "best" population is included in the subset with a given level of confidence. The parameters of interest are

$$\mu_i - \max_{j \neq i} \mu_j, \text{ for } i = 1, 2, \ldots, t$$

where $\max_{j \neq i} \mu_j$ is the maximum treatment mean not including μ_i. If $\mu_i - \max_{j \neq i} \mu_j > 0$ then treatment i is the best. On the other hand, if $\mu_i - \max_{j \neq i} \mu_j < 0$, then treatment i is not the best.

MCB simultaneous confidence intervals (SCI) for $\mu_i - \max_{j \neq i} \mu_j$ are constrained to include 0 with the view that no two treatments ever have identical long run averages (Hsu, 1996). The *constrained* MCB confidence interval asserts treatment i is one of the best if the interval for $\mu_i - \max_{j \neq i} \mu_j$ *includes* 0 or has a *lower* bound of 0. Conversely, if the *upper* interval bound for $\mu_i - \max_{j \neq i} \mu_j$ is 0, then treatment i is not one of the best. The MCB procedure is described in Display 3.9.

Display 3.9 Multiple Comparisons with the Best Procedure

Calculate the difference, D_i, between each treatment mean, $\overline{y}_i$, and the largest treatment mean of the remaining treatments, $\max_{j \neq i} (\overline{y}_j)$, as

$$D_i = \overline{y}_i - \max_{j \neq i} (\overline{y}_j), \text{ for } i = 1, 2, \ldots, t \qquad \textbf{(3.30)}$$

and the quantity M

$$M = d_{\alpha,k,\nu} \sqrt{\frac{2s^2}{r}} \qquad \textbf{(3.31)}$$

where $d_{\alpha,k,\nu}$ is the tabled statistic for one-sided comparisons in Appendix Table VI for an experimental error rate of α_e, $k = t - 1$ comparisons, and ν degrees of freedom for the experimental variance, $s^2 = MSE$.

100(1 − α)% Simultaneous Constrained Confidence Intervals

The lower confidence bound for $\mu_i - \max_{j \neq i} (\overline{y}_j)$ is

$$L = \begin{cases} D_i - M & \text{if } (D_i - M) < 0 \\ 0 & \text{otherwise} \end{cases}$$

and the upper confidence bound for $\mu_i - \max_{j \neq i} (\overline{y}_j)$ is

$$U = \begin{cases} D_i + M & \text{if } (D_i + M) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Example 3.2 Flow Rates Through Filters

The MCB procedure is illustrated using an experiment conducted to evaluate filters with different filtering configurations. The filters were all designed to screen particles above a certain size. The investigators wanted to know which, if any, of the filters allowed the highest flow rate of a particle slurry under constant pressure.

They had developed 6 filter configurations for testing purposes. Four replicate filters of each configuration were constructed for the experiment. The 24 filters were tested in random order for a completely randomized experiment design. The average flow rates for the 6 filter types were

Filter:	A	B	C	D	E	F
Mean:	8.29	7.23	7.54	8.10	8.59	7.10

The estimate of experimental error variance for the experiment was $MSE = 0.08$ with 18 degrees of freedom.

MCB Simultaneous Confidence Intervals

We can calculate the 95% confidence interval for a comparison of filter B with the best of the other filters to illustrate the procedure. The mean for filter B is $\overline{y}_2 = 7.23$, and filter E has the largest mean among all the remaining filters, so that $\max_{j \neq 2} (\overline{y}_j) = \overline{y}_5 = 8.59$. Then $D_2 = 7.23 - 8.59 = -1.36$.

The value for $d_{\alpha,k,\nu}$ in Equation (3.31) is found from Appendix Table VI with $k = 5$, $\alpha_E = .05$, and $\nu = 18$ degrees of freedom for $MSE = 0.08$. The appropriate value is $d_{.05,5,18} = 2.41$, so that with $r = 4$ replications and $MSE = 0.08$

$$M = 2.41\sqrt{\frac{2(0.08)}{4}} = 0.48$$

The required quantities are $D_2 - M = -1.36 - 0.48 = -1.84$ and $D_2 + M = -1.36 + .48 = -0.88$. Using the rules for upper and lower limits in Display 3.9, $L = -1.84$ because $D_2 - M < 0$ and $U = 0$ because $D_2 + M$ is not greater than 0. The upper and lower bounds are shown for each of the comparisons in Table 3.9.

Four of the filters (B, C, D, and F) have upper bounds of 0 and thus are not the "best" filters. Of the remaining two filters (A and E) neither is clearly the best, since both intervals include 0, which in turn implies that each is one of the best filters with 95% confidence. In addition, their lower bounds, -0.78 for A and -0.18 for E, are also close to 0 relative to the lower bounds for the other filters. A lower bound close to 0 indicates the treatment is close to the best (Hsu, 1996). Note the SCI not only provide the means to identify the best treatment(s) but also give information about how far removed each of the treatments is from the best. Based on the lower bound of the intervals in Table 3.9 it is easy to see that filters B, C, and F in particular are the most removed from the best treatment.

If a treatment is clearly the only best, then the lower bound will be 0. To illustrate suppose filter E had a mean of $\overline{y}_5 = 9.00$ rather than 8.59. The next largest mean is filter A with $\overline{y}_1 = 8.29$. The constrained confidence interval comparing filter E with filter A would be

Table 3.9 MCB procedure with flow rate means of six filter types

Filter	$\overline{y}_i$	$\max_{j \neq i} (\overline{y}_j)$	D_i	$D_i - M$	$D_i + M$	*95% SCI* (L, U)	*Select?**
A	8.29	8.59	−0.30	−0.78	0.18	(−0.78, 0.18)	Yes
B	7.23	8.59	−1.36	−1.84	−0.88	(−1.84, 0)	No
C	7.54	8.59	−1.05	−1.53	−0.57	(−1.53, 0)	No
D	8.10	8.59	−0.49	−0.97	−0.01	(−0.97, 0)	No
E	8.59	8.29	0.30	−0.18	0.78	(−0.18, 0.78)	Yes
F	7.10	8.59	−1.49	−1.97	−1.01	(−1.97, 0)	No

Select as "best" when $D_i + M > 0$.

$$D_5 - M = 9.00 - 8.29 - 0.48 = 0.23$$
$$D_5 + M = 9.00 - 8.29 + 0.48 = 1.19$$

so that with $D_5 - M > 0$ and $D_5 + M > 0$ the constrained interval bounds are (0, 1.19), and the interval with a lower bound of 0 indicates filter E is the best. Likewise, filter A would not be one of the best under these circumstances since

$$D_1 - M = 8.29 - 9.00 - 0.48 = -1.19$$
$$D_1 + M = 8.29 - 9.00 + 0.48 = -0.23$$

so that with $D_1 - M < 0$ and $D_1 + M < 0$ the constrained interval bounds are $(-1.19, 0)$. Thus, with an upper bound of 0 filter A would not be the best treatment or among the set of best treatments.

Selecting the Subset with the Largest Mean

If the only interest is which treatment or treatments constitute the "best" without regard to how much their effects may differ from the others, then a simple selection rule can be used (Hsu, 1984). The MCB rule selects a treatment into the best subset with a probability of correct selection, $P(\text{CS}) = 1 - \alpha$, if

$$D_i + M > 0 \tag{3.32}$$

The MCB procedure to select the best filter types with $P(\text{CS}) = 0.95$ uses the $D_i + M$ column information given in Table 3.9. The treatments in the best subset are those for which $D_i + M > 0$. Only filters A and E have values of $D_i + M > 0, 0.18$ and 0.78 respectively. The best subset includes these two filter types with a probability $P(\text{CS}) = 0.95$ that the best filter is in the subset A and E.

The conclusions from the subset selection procedure do not differ from the SCI results. Both filter types A and E were selected to the best subset. However, the SCI are more informative because they indicated how close to 0 the lower bounds were

for each of the filter types, which in turn gives the investigator more information regarding flow rate performance for each of the filters relative to that for the filters with the highest flow rates. The subset selection procedure limits the information to whether a filter is in the best subset and indicates no more than that.

Multiple Comparisons with the Smallest Mean

In some studies the treatment with the smallest mean is the "best," such as in the meat storage experiment in which a packaging would be best if it had fewer bacteria on the surface. Multiple comparisons with the smallest mean can be made with simple modifications of the rule for selecting the largest mean in Display 3.9.

Calculate the difference, D_i, between each treatment mean, $\overline{y}_i$, and the smallest mean of the remaining treatments, $\min_{j \neq i} (\overline{y}_j)$, as

$$D_i = \overline{y}_i - \min_{j \neq i} (\overline{y}_j) \text{ for } i = 1, 2, \ldots, t \qquad \textbf{(3.33)}$$

The $100(1-\alpha)\%$ simultaneous confidence interval lower (L) and upper (U) limits are found as follows: The lower confidence bound for $\mu_i - \min_{j \neq i} \mu_j$ is

$$L = \begin{cases} D_i - M & \text{if } (D_i - M) < 0 \\ 0 & \text{otherwise} \end{cases}$$

and the upper confidence bound for $\mu_i - \min_{j \neq i} \mu_i$ is

$$U = \begin{cases} D_i + M & \text{if } (D_i + M) > 0 \\ 0 & \text{otherwise} \end{cases}$$

The constrained MCB confidence intervals for comparisons with the smallest treatment mean are interpreted just opposite of those for comparisons with the largest treatment mean. If the interval for $\mu_i - \min_{j \neq i} \mu_j$ *includes* 0 or has an *upper* bound of 0, then treatment i is the best treatment. Conversely, if the *lower* bound is 0, then treatment i is not the best treatment.

Multiple comparisons with the best as the treatment with the smallest mean is illustrated with the meat storage study and an objective to select the packaging material with the least amount of bacterial growth. The experimental error variance for the experiment was $MSE = 0.116$ with 8 degrees of freedom, and there were $r = 3$ replications. The required statistics for the critical value of M in Equation (3.31) are $d_{.05,3,8} = 2.42$ and $\sqrt{2MSE/r} = 0.278$, so that

$$M = d_{.05,3,8}\sqrt{\frac{2MSE}{r}} = 2.42(0.278) = 0.67$$

The treatment means and 95% simultaneous confidence interval estimates for comparisons with the smallest mean are shown in Table 3.10. The procedure is illustrated for the commercial wrap treatment with mean $\overline{y}_1 = 7.48$ and the

Table 3.10 Selection of the Treatment Subset with the Smallest Means in the Meat Storage Experiment

Treatment	$\overline{y}_i$	$\min_{j \neq i} (\overline{y}_j)$	D_i	$D_i - M$	$D_i + M$	*95% SCI* (L, U)	*Select?**
Commercial	7.48	3.36	4.12	3.45	4.79	(0, 4.79)	No
Vacuum	5.50	3.36	2.14	1.47	2.81	(0, 2.81)	No
Mixed	7.26	3.36	3.90	3.23	4.57	(0, 4.57)	No
Pure CO_2	3.36	5.50	−2.14	−2.81	−1.47	(−2.81, 0)	Yes

* Select as "best" when $D_i - M < 0$.

treatment with the minimum mean, pure CO_2, so that $\min_{j \neq 1} (\overline{y}_j) = \overline{y}_4 = 3.36$ to give $D_1 = 7.48 - 3.36 = 4.12$. The lower limit is $L = 0$ since $D_1 - M = 4.12 - 0.67 = 3.45$ is not less than 0, and the upper limit is $U = 4.79$ since $D_1 + M = 4.12 + 0.67 = 4.79$ is greater than 0. With a lower bound of 0 we can assert that the commercial wrap is not the best treatment. The best treatment is pure CO_2 since it is the only treatment in Table 3.10 with an upper confidence interval bound of 0. The upper bounds for all other treatments are considerably removed from 0 and clearly cannot be considered close to the best.

Selecting the Subset with the Smallest Mean

Again, if the only interest is which treatment or treatments constitute the best without regard to how much their effects may differ from the others, then a simple selection rule can be used (Hsu, 1984). The MCB rule selects a treatment into the best subset with a probability of correct selection, $P(\text{CS}) = 1 - \alpha$, if

$$D_i - M < 0 \qquad \textbf{(3.34)}$$

The MCB procedure to select the best meat packaging with $P(\text{CS}) = 0.95$ uses the $D_i - M$ column information given in Table 3.10. The treatments in the best subset are those for which $D_i - M < 0$. The only treatment mean with $D_i - M < 0$ is that for pure CO_2. Pure CO_2 is selected as the treatment with the lowest bacterial growth with a probability of correct selection $P(\text{CS}) = .95$.

Unequal Replications and Complex Models

Hsu (1996) indicates that critical values must be computed separately for each comparison with unequal replication numbers in the completely randomized design. Some computer programs have incorporated routines to compute the simultaneous constrained confidence intervals with unequal replication numbers. No approximations based on probability inequalities were given by Hsu (1996).

If a more complex blocking or treatment design is used and all differences, $\mu_i - \mu_j$, have the same variance—that is, the design is *variance balanced*—then the standard MCB procedure in this section may be used.

If the variances for differences, $\mu_i - \mu_j$, are not all the same in more complex designs, then Hsu (1996) recommends several approximations based on probability inequalities. The approximation based on the *Bonferroni inequality* uses the Bonferroni t for k comparisons with ν degrees of freedom at the appropriate level of α in place of $d_{\alpha,k,\nu}$. The least squares estimate, $\widehat{\mu}_i - \widehat{\mu}_j$, and its standard error should be used; they can be obtained from most computer programs.

3.7 Comparison of All Treatments with a Control

In many studies one of the treatments acts as a control treatment for some or all of the remaining treatments. (Different types of control treatments were discussed in Section 1.4.) Determining whether the mean responses for the treatments differ from that for the control treatment is sometimes of interest. Dunnett (1955) introduced a procedure for this purpose based on an experimentwise error rate.

Simultaneous Confidence Intervals

The tabled statistic to compute $100(1-\alpha)\%$ simultaneous confidence intervals for differences between the individual treatment means and the control mean $\mu_i - \mu_c$ using the Dunnett procedure is based on the same statistic used for the multiple comparisons with the best procedure in Section 3.6. The Dunnett test to compare each treatment mean, $\overline{y}_i$, with the control treatment mean $\overline{y}_c$ is described in Display 3.10.

Suppose filter F in Example 3.2 serves as a control. The five 95% SCI comparisons for control versus treatment means are shown in Table 3.11. The mean differences between F and the $k = 5$ other filters appear in the third column. The critical value of the Dunnett statistic for a two-sided test with an error rate of $\alpha_E = .05$ is $d_{.05,5,18} = 2.76$. The standard error of the difference is $\sqrt{2MSE/r} = \sqrt{2(0.08)/4} = 0.2$. The Dunnett criterion is

$$D(5, .05) = 2.76(0.2) = 0.55$$

An example calculation is illustrated with filter A. The lower limit is

$$L = \overline{y}_1 - \overline{y}_c - D(5, .05) = 8.29 - 7.10 - 0.55 = 0.64$$

and the upper limit is

$$U = \overline{y}_1 - \overline{y}_c + D(5, .05) = 8.29 - 7.10 + 0.55 = 1.74$$

Filter A is superior to the control filter, F, since the lower bound of the interval is greater than 0. Based on the interval estimates filters A, D, and E are superior to filter F. Also, note that filter E is the most removed from filter F since its lower bound is greater than the lower bound of either filter A or D. Filters B and C do not differ from the control since the intervals include 0.

Display 3.10 The Dunnett Method for a Comparison of All Treatments with a Control

$100(1-\alpha)\%$ Simultaneous Confidence Intervals for $\mu_i - \mu_c$

The Dunnett criterion to compare k treatments to the control is

$$D(k, \alpha_E) = d_{\alpha,k,\nu}\sqrt{\frac{2s^2}{r}} \tag{3.35}$$

Simultaneous two-sided confidence interval estimates for the differences between the individual treatment means and the control means $\mu_i - \mu_c$ are

$$\overline{y}_i - \overline{y}_c \pm D(k, \alpha_E) \tag{3.36}$$

One-sided interval lower bounds if superiority is manifested by a treatment mean *greater* than the control mean are

$$\overline{y}_i - \overline{y}_c - D(k, \alpha_E) \tag{3.37}$$

One-sided interval upper bounds if superiority is manifested by a treatment mean *less* than the control mean are

$$\overline{y}_i - \overline{y}_c + D(k, \alpha_E) \tag{3.38}$$

The values of $d_{\alpha,k,\nu}$ for the two-sided or one-sided Dunnett method are found in Appendix Table VI for k treatments, an experimentwise Type I error of α_E, and ν degrees of freedom for the estimate of experimental error variance.

Table 3.11 Results of the Dunnett test comparing the mean of the control filter, F, with that of all other filters

Filter	*Mean*	$\overline{y}_i - \overline{y}_c$	*95% SCI* (L, U)	$\lvert\overline{y}_i - \overline{y}_c\rvert$	*Different from control?**
F	$\overline{y}_c = 7.10$	—	–	–	—
A	8.29	1.19	(0.64, 1.74)	1.19	Yes
B	7.23	0.13	(− 0.42, 0.68)	0.13	No
C	7.54	0.44	(− 0.11, 0.99)	0.44	No
D	8.10	1.00	(0.45, 1.55)	1.00	Yes
E	8.59	1.49	(0.94, 2.04)	1.49	Yes

*If $\lvert\overline{y}_i - \overline{y}_c\rvert$ exceeds $D(5, .05) = 0.55$, then the filter mean is different from that of filter F (control).

Testing Hypotheses About $\mu_i - \mu_c$

If the investigator only wants to know if a treatment mean is significantly different from the control mean with significance level α, then the following two-sided or

one-sided tests can be conducted: For the two-sided alternative with H_0: $\mu_i = \mu_c$ versus H_a: $\mu_i \neq \mu_c$, reject the null hypothesis if

$$|\overline{y}_i - \overline{y}_c| > D(k, \alpha_E) \tag{3.39}$$

For the one-sided alternative with H_0: $\mu_i \leq \mu_c$ versus H_a: $\mu_i > \mu_c$, reject the null hypothesis if

$$(\overline{y}_i - \overline{y}_c) > D(k, \alpha_E) \tag{3.40}$$

For the one-sided alternative with H_0: $\mu_i \geq \mu_c$ versus H_a: $\mu_i < \mu_c$; reject the null hypothesis if

$$(\overline{y}_i - \overline{y}_c) < -D(k, \alpha_E) \tag{3.41}$$

Results of the test with the two-sided alternatives are shown in Table 3.11. The differences for filters A, D, and E exceed $D(5, .05) = 0.55$, and their flow rates differ significantly from that for filter F. The two-sided tests exemplify a confident inequalities inference. We only state with error rate .05 that filters A, D, and E differ from filter F, and filters B and C do not differ from filter F. One can only deduce indirectly from the size and sign which differences were positive and which were the largest as opposed to the SCI, which directly give us information about size and direction.

Unequal Replications and Complex Models

Hsu (1996) indicates that critical values must be computed separately for each comparison with unequal replication numbers in the completely randomized design. Some computer programs have incorporated routines to compute the simultaneous constrained confidence intervals with unequal replication numbers. An approximation based on the Bonferroni inequality can be used that substitutes the Bonferroni t for k comparisons with ν degrees of freedom at the appropriate level of α in place of $d_{\alpha,k,\nu}$ and uses $\sqrt{2s^2[1/r_i + 1/r_c]}$ for the standard error of the difference.

If a more complex blocking or treatment design is used and all differences, $\mu_i - \mu_j$, have the same variance—that is, the design is *variance balanced*—then the standard procedure in this section may be used.

If the variances for differences, $\mu_i - \mu_c$, are not all the same in more complex designs, then Hsu (1996) presents several approximations, some of which are based on probability inequalities. The approximation based on the Bonferroni inequality uses the Bonferroni t for k comparisons with ν degrees of freedom at the appropriate level of α in place of $d_{\alpha,k,\nu}$. The least squares estimate, $\widehat{\mu}_i - \widehat{\mu}_c$, and its standard error should be used; they can be obtained from most computer programs.

Dunnett (1964) provided adjustments to the critical values in Appendix Table VI for unequal replication numbers. A conservative upper bound given by Fleiss (1986) for values in Appendix Table VI is $md_{\alpha,k,\nu}$, where

$$m \leq 1 + 0.07\left(1 - \frac{r_i}{r_c}\right) \tag{3.42}$$

Dunnett has shown the optimal ratio of replication numbers is r/r_c where r is the common replication number for each treatment and r_c is the replication number for the control treatment.

3.8 Pairwise Comparison of All Treatments

Some investigators compare each treatment mean with each of the other treatment means using **pairwise comparisons**. The parameters of interest are all pairwise differences among the treatment means, $\mu_i - \mu_j$ for all $i \neq j$, resulting in $t(t-1)/2$ comparisons. Most frequently, applications of these methods have an objective to detect significant inequalities, $\mu_i \neq \mu_j$ for all $i \neq j$.

The indiscriminate use of pairwise comparison procedures in this manner for the analysis of experimental results can lead to the tendency to place a reliance on statistical significance alone to drive the inferential procedure in data analysis. It is therefore possible to lose sight of the research objectives, and the investigator's focus may diverge from the pursuit of biological or physical understanding to that of statistical significance.

Ideally, investigators would want to make the comparisons with an experimentwise error rate in the strong sense as they were made with multiple comparisons with the best (Section 3.6) and multiple comparisons with the control (Section 3.7). Several methods for pairwise comparisons will be presented in this section with some discussion of their properties.

The Tukey Method

A procedure providing an experimentwise rate in the strong sense was developed by Tukey (1949a) for pairwise comparison of all treatment means and is used to obtain $100(1-\alpha)\%$ simultaneous confidence intervals. The test has been called by various names, including the Honestly Significant Difference. The Tukey method is described in Display 3.11.

The Tukey method is based on the Studentized range statistic

$$q = \frac{\overline{y}(\text{largest}) - \overline{y}(\text{smallest})}{\sqrt{\frac{s^2}{r}}} \tag{3.43}$$

where $\overline{y}$(largest) is the largest mean in an ordered group of means in an experiment and $\overline{y}$(smallest) is the smallest of the means. The difference or range is divided by the standard error of a treatment mean, from which the statistic derives the name of Studentized range statistic.

Display 3.11 The Tukey Method for All Pairwise Comparisons

For a group of k treatment means compute the Honestly Significant Difference as

$$\text{HSD}(k, \alpha_E) = q_{\alpha,k,\nu}\sqrt{\frac{s^2}{r}} \tag{3.44}$$

where $q_{\alpha,k,\nu}$ is the Studentized range statistic for a range of k treatment means in an ordered array. Critical values for an experimentwise error rate, α_E, and ν degrees of freedom can be found in Appendix Table VII.

100(1 − α)% Simultaneous Confidence Intervals

Simultaneous two-sided interval estimates for the absolute value of all pairwise differences, $\mu_i - \mu_j$ for all $i < j$, are

$$|\overline{y}_i - \overline{y}_j| \pm \text{HSD}(k, \alpha_E) \tag{3.45}$$

100(1 − α)% Confident Inequalities Test

Two treatment means are declared not equal, $\mu_i - \mu_j \neq 0$, if

$$|\,\overline{y}_i - \overline{y}_j\,| > \text{HSD}(k, \alpha_E) \tag{3.46}$$

The absolute difference $|\overline{y}_i - \overline{y}_j|$ is given in Display 3.11 for the confidence intervals because the location of the two means in the calculated difference, $\overline{y}_i - \overline{y}_j$, is arbitrary, with the sign of the difference depending on whether one calculates $\overline{y}_i - \overline{y}_j$ or $\overline{y}_j - \overline{y}_i$. Thus, the absolute difference is equivalent to always subtracting the smaller mean from the larger. If the direction of a particular difference is necessary, then calculate the interval with the sign of the difference considered. The particulars of the study will dictate whether absolute or signed differences are best used for specific comparisons.

Example 3.3 Strength of Welds

Pairwise comparison tests are illustrated using an experiment conducted to compare the strength of welds produced by four different welding techniques. Each welding technique was used to weld five pairs of metal plates in a completely randomized design. The average strengths for the five welds of each technique were

Technique:	**A**	**B**	**C**	**D**
Mean:	69	83	75	71

The estimate of experimental error variance for the experiment was $MSE = 15$ with 16 degrees of freedom.

Applying the Tukey method to the data of Example 3.3, the Studentized range statistic with an experimentwise error rate of $\alpha_E = .05$ is found from Appendix Table VII as $q_{.05,4,16} = 4.05$, where there are $k = 4$ treatment means in the ordered array and $\nu = 16$ degrees of freedom for $MSE = s^2$. The standard error is $\sqrt{MSE/r} = \sqrt{15/5} = 1.73$. The computed HSD is HSD(4, .05) = 4.05(1.73) = 7.0. The 95% SCI and results of the confident inequalities test are shown in Table 3.12.

Table 3.12 Results of the Tukey method for differences between weld strength means for four welding methods

Comparison	$\lvert\overline{y}_i - \overline{y}_j\rvert$	*95% SCI* (L, U)	*Different from 0?**
A vs. B	14	(7, 21)	Yes
A vs. C	6	(−1, 13)	No
A vs. D	2	(−5, 9)	No
B vs. C	8	(1, 15)	Yes
B vs. D	12	(5, 19)	Yes
C vs. D	4	(−3, 11)	No

* The absolute difference exceeds HSD(.05) = 7.0.

Two treatment means are different with 95% confidence if the 95% SCI interval does not include 0. Method B is significantly different from all other methods, but no other treatments differ from one another. The amount method B differs from the other three methods can be assessed by the magnitude of the lower bound. Method B differs most from method A and least from method C.

Inference on the basis of confidence inequalities declares two treatment means different if the absolute difference, $|\,\overline{y}_i - \overline{y}_j\,|$, exceeds HSD(4, .05) = 7.0. The HSD test judges the weld strengths of method B to be different from the weld strengths of all other methods, but no other differences are significant. However, inference about the direction and magnitude of the inequalities is not possible.

Unequal Replications and Complex Models with the Tukey Method

For unequal replications in the completely randomized design, Tukey in 1953 (see Tukey, 1994) and Kramer (1956) proposed approximate simultaneous confidence intervals, which substitute

$$\sqrt{\frac{s^2}{2}\left(\frac{1}{r_i} + \frac{1}{r_j}\right)} \tag{3.47}$$

for $\sqrt{s^2/r}$ in Equation (3.43). The approximation has become known as the Tukey–Kramer approximation, which Hayter (1984) showed to be conservative.

If the variances for differences, $\mu_i - \mu_j$, are not all the same in more complex designs, then the least squares estimates, $\widehat{\mu}_i - \widehat{\mu}_j$, and their variances, $s^2_{(\widehat{\mu}_i - \widehat{\mu}_j)}$, are used for an approximation to the exact form. The simultaneous intervals are

$$|\widehat{\mu}_i - \widehat{\mu}_j| \pm q_{\alpha,k,\nu}\sqrt{\frac{1}{2}s^2_{(\widehat{\mu}_i - \widehat{\mu}_k)}} \qquad \textbf{(3.48)}$$

If a more complex blocking or treatment design is used and all differences, $\mu_i - \mu_j$, have the same variance—that is, the design is *variance balanced*—then the standard procedure in this section may be used, using the least squares estimates, $\widehat{\mu}_i - \widehat{\mu}_j$, and their common variance of the difference, $s^2_{(\widehat{\mu}_i - \widehat{\mu}_j)}$ in Equation (3.48). The strong sense experimentwise error rate will be exact in this case.

Tests of Homogeneity, $\mu_1 = \mu_2 = \cdots = \mu_t$

Many of the popular pairwise multiple comparison tests have been used with experimentwise error rates evaluated under a restricted assumption that all treatment means are equal, or $\mu_1 = \mu_2 = \cdots = \mu_t$, thus providing experimentwise error rates in the *weak* sense. Some of them use a single criterion for declaration of significance. Other tests are referred to as *multiple-range tests* since they use multiple criteria for declaration of significance, where the value of a criterion for one comparison depends upon how far apart the two means are in the ordered array of all the treatment means. The *Least Significant Difference, LSD(α),* is the most common test of homogeneity that uses a single criterion. The *Student-Newman-Keuls, SNK(k,α),* is an example of a multiple-range test that is a test of homogeneity. Both tests are illustrated below.

The Least Significant Difference (LSD)

Each hypothesis H_0: $\mu_i = \mu_j$ versus H_a: $\mu_i \neq \mu_j$ can be tested with the Student t statistic:

$$t_0 = \frac{\overline{y}_i - \overline{y}_j}{\sqrt{s^2\left[\frac{1}{r_i} + \frac{1}{r_j}\right]}} \quad \text{for all } i \neq j \qquad \textbf{(3.49)}$$

When the Type I error probability is set at some value α and the variance s^2 has ν degrees of freedom, the null hypothesis is rejected for any observed value of $|\overline{y}_i - \overline{y}_j|$ such that $|t_0| > t_{\alpha/2,\nu}$. The **Least Significant Difference, LSD(α)**, is an abbreviated method of conducting all possible pairwise t tests as shown in Display 3.12.

Display 3.12 Least Significant Difference for All Pairwise Comparisons

For any pair of observed treatment means, $\overline{y}_i$ and $\overline{y}_j$, the Least Significant Difference is

$$\text{LSD}(\alpha) = t_{\alpha/2,\nu}\sqrt{s^2\left[\frac{1}{r_i} + \frac{1}{r_j}\right]} \qquad \textbf{(3.50)}$$

The null hypothesis H_0: $\mu_i = \mu_j$ is rejected if

$$|\overline{y}_i - \overline{y}_j| > \text{LSD}(\alpha)$$

A modification by Fisher (1960) controls the weak sense experimentwise error rate. The LSD is used to test pairwise comparisons only if the null hypothesis is rejected in the analysis of variance F test. If the null hypothesis is not rejected on the basis of the F test, then all treatment means are assumed to be the same and no further testing is done. The procedure often is referred to as the protected LSD. Carmer and Swanson (1973) provided empirical demonstration that the experimentwise error rate with the protected LSD was almost the same as the significance level of the F test used as the determinate when α was set at the .05 significance level for the LSD.

Investigations by Finner (1990) and Hayter (1986) showed the LSD test was a confident inequalities method if the number of treatments was less than 3, but that for $t > 3$ it was not a confident inequalities method.

The LSD test is illustrated with the weld data from Example 3.3. The computation of the LSD requires the critical value for the Student t test, $t_{.025,16} = 2.12$, and the standard error of the difference between two treatment means, $\sqrt{2MSE/r} = \sqrt{2(15)/5} = 2.45$. The computed LSD is LSD(.05) = 2.12(2.45) = 5.2.

The null hypothesis H_0: $\mu_i = \mu_j$ is rejected if

$$|\overline{y}_i - \overline{y}_j| > \text{LSD}(.05) = 5.2$$

A convenient method for testing is to form a table of differences with the means ordered from smallest to largest as shown in Table 3.13. The differences in the table are each computed as a difference between a mean of the column head and a mean of lesser value in the leftmost column. An asterisk indicates the differences that exceeded LSD(.05) = 5.2.

The result of all pairwise comparisons in Table 3.13 indicates that the weld strength of method B exceeds that of all other methods, and the weld strength of method C exceeds that of method A.

The Student-Newman-Keuls (SNK) Multiple-Range Test

The SNK test is one of many multiple-range tests. It is based on the Studentized range statistic in Equation (3.43), but in contrast to the Tukey method it results in a

Table 3.13 Results of the LSD test on differences between the weld strength means for four welding methods

		Method			
		A	*D*	*C*	*B*
Method	*Mean*	69	71	75	83
A	69	–	2	6*	14*
D	71		–	4	12*
C	75			–	8*
B	83				–

* The difference exceeds LSD(.05) = 5.2.

homogeneity test with experimentwise error rates in the weak sense (Hsu, 1996). The critical value for the Studentized range with the SNK test is based on the range of the particular pair of means being tested within the entire set of ordered means. The test was developed independently by Newman (1939) and Keuls (1952) and is categorized as a multiple-range test since two or more ranges among the means are used for the test criteria. The SNK test is described in Display 3.13.

Display 3.13 Student-Newman-Keuls Multiple Range Test

The SNK criterion is

$$\text{SNK}(k, \alpha_E) = q_{\alpha,k,\nu}\sqrt{\frac{s^2}{r}} \quad \text{for } k = 2, 3, \ldots, \tag{3.51}$$

where $q_{\alpha,k,\nu}$ is the Studentized range statistic, k is the number of means in the range, ν is the number of degrees of freedom for the estimate of experimental error variance s^2, and α_E is the experimentwise error rate for a range of k means.

For the largest and smallest means in a range of k means, say $\overline{y}_i$ and $\overline{y}_j$, the null hypothesis H_0: $\mu_i = \mu_j$ is rejected if

$$|\overline{y}_i - \overline{y}_j| > \text{SNK}(k, \alpha_E)$$

The test is not conducted if there is a range of means of size greater than k containing $\overline{y}_i$ and $\overline{y}_j$ that is not significant by the SNK criteria.

The SNK test is presented here to demonstrate the methods for multiple-range tests and also because tables of critical values for the Studentized range statistic for a constant α are readily available. A multiple-range test that provides a much stronger inference with confident inequalities is presented after the SNK test. Although the other test also uses the Studentized range statistic it has not become popular, probably because its critical values are not as easily accessible as those for the SNK test.

The treatment means are ordered from smallest to largest

$$\overline{y}_{[1]} \le \overline{y}_{[2]} \le \overline{y}_{[3]} \le \cdots \le \overline{y}_{[t]}$$

where $\overline{y}_{[1]}$ is the treatment mean with the smallest value and $\overline{y}_{[t]}$ is the treatment mean with the largest value. The critical value for each pair of means depends on the number of means in the range of the particular pair of means under test.

The SNK test of the means in Example 3.3 with an experimentwise Type I error rate of $\alpha_E = .05$ requires three critical values of the Studentized range statistic from Appendix Table VII, with $\alpha_E = .05$, $\nu = 16$, and $k = 2$, 3, and 4. With standard error $\sqrt{MSE/r} = \sqrt{15/5} = 1.73$, the SNK statistic is computed from

$$\text{SNK}(k, .05) = q_{.05,k,16}(1.73) \qquad \text{for } k = 2, 3, 4$$

The three critical values of the Studentized range statistic and SNK(k, .05) for Example 3.3 are

k:	2	3	4
$q_{.05,k,16}$:	3.00	3.65	4.05
SNK(k, .05):	5.2	6.3	7.0

The critical values for the SNK test increase as the number of means in a range increases. As the distance between two means in the ordered array increases, a larger difference between the means is required to declare the means different from one another. For the minimum range of $k = 2$ the SNK(2, .05) is equal to the LSD(.05) = 5.2, and for the maximum range of $k = 4$ the SNK(4, .05) is equal to the HSD(4, .05) = 7.0. The SNK test is more conservative than the LSD but less conservative than the HSD in the required differences for rejection of the null hypothesis. The results of the SNK(k, .05) test for Example 3.3 are shown in Table 3.14.

Table 3.14 Results of the SNK(k, .05) on differences between the weld strength means for four welding methods

		Method					
		A	*D*	*C*	*B*		
Method	*Mean*	69	71	75	83	k	*SNK*(k, .05)
A	69	–	2	6	14*	4	7.0
D	71		–	4	12*	3	6.3
C	75			–	8*	2	5.2
B	83						

*Differences shown are a larger mean minus a smaller mean. The difference exceeds SNK(k, .05).

The differences between treatment means for the same range, k, are found on a diagonal running from the upper left to the lower right in Table 3.14. The SNK test commences with a comparison between the minimum and maximum means, $k = 4$. If the maximum difference does not exceed the critical value no further testing is

conducted. If the maximum difference is significant the differences among the means with range $(k-1)$ are tested.

If any pair of means with range $(k-1)$ are not significant no further testing is conducted for any other pairs of means between that specific pair of means. By definition, no subgroup of means contained in a nonsignificant group of means can be significant. For example, if there is no significant difference between $\overline{y}_{[1]}$ and $\overline{y}_{[3]}$, then no test should be performed for $\overline{y}_{[1]}$ versus $\overline{y}_{[2]}$ or $\overline{y}_{[2]}$ versus $\overline{y}_{[3]}$. The test proceeds in this manner until there are no further significant ranges.

The maximum difference, $k = 4$, is 14 for B versus A. The difference exceeds SNK(4, .05) = 7.0, so the test proceeds with comparisons for a range of $k = 3$, B versus D and C versus A. The difference for C versus A does not exceed SNK(3, .05) = 6.3. Therefore, no more tests are conducted between pairs of means in the range C to A. They are D versus A and C. The difference for B versus D exceeds SNK(3, .05) = 6.3, so testing continues within the groups of means between B and D. The difference for B versus C exceeds SNK(2, .05) = 5.2 and is the only test necessary with a range of $k = 2$ means. The SNK test judges method B to be different from all other methods with no differences among the other methods.

With unequal replication numbers for the HSD and SNK tests the harmonic mean of the replication numbers from all treatment groups, r_h, is often used in the standard error estimate for all comparisons to simplify the calculations. However, Hsu (1996) showed its use led to invalid statistical inference in general by reducing the confidence levels considerably. The harmonic mean r_h is

$$r_h = \left[\frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{r_i}\right)\right]^{-1} \tag{3.52}$$

Multiple Range Tests for Confident Inequalities

Several multiple-range tests have been developed that provide confident inequalities inference, which is a stronger inference than that provided by the SNK test. Einot and Gabriel (1975) proposed a choice of α_k to test the difference between two means with a range of k in the set of t treatment means. A modification suggested by a number of authors has led to common use as

$$\alpha_k = \begin{cases} 1-(1-\alpha_E)^{\frac{k}{t}} & \text{if } k = 2, \ldots, k-2 \\ \alpha & \text{if } k = t-1, t \end{cases} \tag{3.53}$$

for the desired experimentwise error rate α_E. The Studentized range statistic for a range of k means would be used for the critical value as $q_{\alpha_k,k,\nu}$.

The test will provide confident inequalities inference if the critical values are nondecreasing with increasing values of k for the range (Hsu, 1996). The test can be conducted fairly easily if a computer program is available to compute quantiles for the Studentized range statistic.

The values of α_k and critical values of the Studentized range statistic calculated from a computer program for the weld strength example are shown below and can be compared with those for the SNK test given earlier. The statistic for a decision regarding significance of a difference between two means is $\text{EG}(k, \alpha_k) = q_{\alpha_k,k,\nu}\sqrt{s^2/r}$. Recall there were $t = 4$ treatments, $r = 5$ replications, and a standard error of 1.73 with $\nu = 16$ degrees of freedom for the error mean square.

k:	2	3	4
α_k:	.025	.05	.05
$q_{\alpha_k,k,16}$:	3.49	3.65	4.05
EG(k, .05):	6.0	6.3	7.0

Notice the values of the Studentized range statistic for $k = 3$ and 4 are the same as those for the SNK test, but the value for $k = 2$ is somewhat larger because of the different choice for α_k. The SNK test uses $\alpha_k = \alpha_E$ for all values of k. The significant comparisons using EG(k, α_k) are the same as those with the SNK test; however, the inference strength has increased to that of confident inequalities from strict homogeneity (which the SNK test provided). The differences between the two tests become more pronounced as the number of treatments increase.

3.9 Summary Comments on Multiple Comparisons

Research hypotheses and treatment designs are the engines that drive the methods for analysis of observed results from a study. A set of treatments structured to address certain research hypotheses lead naturally to planned comparisons, such as a set of orthogonal (or nonorthogonal) contrasts, polynomial regressions, and comparisons of all treatments with a control.

The MCB Procedure for Screening Studies

A set of unstructured treatments challenges the investigator to select an appropriate protocol for decision making. Such studies include experiments to evaluate sets of crop cultivars, industrial products, pesticides, and so forth. The MCB procedure to select a subset of treatments with the desired response is the logical choice if the objective is to screen for the best products, pesticides, or cultivars. Subsequent studies may provide the possibility for more structured research hypotheses and treatment designs with opportunities for hypothesis testing.

Multiple Contrasts Require Decisions About Error Rates

Multiple contrasts require the investigator to make some decision relevant to error rates and power of the tests. Comparisonwise error rates are applicable if individual comparisons are the conceptual units of interest. Experimentwise error rates are appropriate if a family of comparisons is the conceptual unit of interest, and the

investigator wants to reduce the chance of too many incorrect decisions in the family of tests.

Probabilities for Type II errors and the power of a test can be determined on an experimentwise basis, but they are usually expressed as comparisonwise rates. The more liberal comparisonwise error rate will result in a lower Type II error rate and a more powerful test given all other conditions constant. The more conservative investigator will utilize the strong sense experimentwise error rate for a family of comparisons. The family of comparisons as a conceptual unit is important to the investigator under these circumstances. A typical family would be the comparison of all treatments with a control using the Dunnett procedure.

Most good experiments are designed to be efficient in the use of existing resources and time, and it is more efficient to design an experiment that answers multiple, related questions. A family of related comparisons will exist within any well-planned experiment, and one comparison cannot be considered in total isolation from all other comparisons; thus, tests based on experimentwise error rates are most appropriate.

Choose a Pairwise Comparison Procedure Consistent with Your Philosophy

The selection of an appropriate pairwise comparison method is difficult since they each have their advantages and disadvantages. The best alternative is to choose a test that is consistent with your philosophy and use it consistently for all of your pairwise comparison tests. A test with a single critical value for each experiment is preferable. Informative discussions on many different pairwise comparison methods can be found in Hsu (1996) and in extensive notes on multiple comparisons that were written by J. W. Tukey in 1953 (Tukey, 1994). Some discussions of their use by Jones (1984), Carmer and Walker (1985), and Saville (1990) may prove useful.

The Tukey method provides the best protection against decision errors, along with the strong inference about magnitude and direction of differences with the $100(1-\alpha)\%$ SCI.

Hayter (1990) provided a method for one-sided simultaneous lower confidence bounds, $\mu_i - \mu_j > L$ for all $i > j$, which is very useful for directional inference in certain types of studies. The bounds are computed as

$$\widehat{\mu}_i - \widehat{\mu}_j - q_\alpha^* \sqrt{\frac{2s^2}{r}} \quad \text{for all } i > j \tag{3.54}$$

These bounds will be much sharper than those supplied by two-sided procedures such as the Tukey method. Tables of critical values for the statistic q_α^* can be found in Hayter and Liu (1996).

A General Recommendation

The general recommendation for conducting multiple comparisons among treatments is to utilize the research hypotheses and treatment design to choose appropriate multiple comparison procedures with the strength of inference desired for the study. Confidence intervals provide the strongest inference with magnitude and direction of differences followed in strength by confident directions and then confident inequalities. A confidence interval procedure that provides magnitude and direction of inference is recommended if pairwise comparisons are your only remaining alternative.

EXERCISES FOR CHAPTER 3

1. Use the data on traffic delay in Exercise 2.1.
   a. Conduct an analysis of variance for the data, and estimate the following contrasts and their standard errors:
      (i) a contrast between the pretimed and the average of the semi- and fully actuated signals
      (ii) a contrast between the semi- and fully actuated signals
   b. Compute the sum of squares of each contrast, and show that their sum is equal to the treatment sum of squares in the analysis of variance.
   c. Test the null hypothesis for each contrast, H_0: $C = 0$, with the Student t test at the .05 level of significance.
   d. Test the null hypothesis in part (c) with the F test at the .05 level of significance.
   e. What is the relationship between the two tests in parts (c) and (d)?

2. Use the data on serum T3 concentrations from the experiments with chickens in Exercise 2.3. Contrasts of interest were the serum T3 concentration differences between successive stages: (1) premolt versus fasting, (2) fasting versus 60 grams of bran, (3) 60 grams of bran versus 80 grams of bran, and (4) 80 grams of bran versus laying mash.
   a. Estimate each of the contrasts and their standard errors.
   b. Test the null hypothesis for one of the contrasts with the Student t test.
   c. Test the null hypothesis for one of the contrasts with the F test.
   d. Suppose you were to test the four contrasts each with a comparisonwise error rate of .05. Compute the maximum experimentwise error rate for this family of four tests.

3. Use the data on lettuce yields in Exercise 2.2.
   a. Compute the analysis of variance for the data.
   b. Determine the best polynomial response function that describes the relationship between lettuce yield and nitrogen fertilizer at the .05 level of significance.
   c. See Table 3.7. Construct a similar table for the results of the current problem.
   d. Plot the observed means along with the estimated equation.

4. Use the data from Exercise 2.1.
   a. Compute the 95% SCI for multiple comparisons with the best with "best" defined as the signal type with the shortest stopped time delay.
   b. Select the signal type(s) with the shortest stopped time delay with a probability of correct selection of .95.

5. A hospital clinical laboratory measures the concentration of cholesterol in patient serum samples with a spectrophotometer. On one particular day the laboratory analyzed samples from eight patients. Two samples from each patient were prepared for analysis. The data that follow are concentrations of cholesterol (mg/dl) .

Patient	*Cholesterol (mg/dl)*
1	167.3, 166.7
2	186.7, 184.2
3	100.0, 107.9
4	214.5, 215.3
5	148.5, 149.5
6	171.5, 167.3
7	161.5, 159.4
8	243.6, 245.5

a. Compute the 95% SCI for multiple comparisons with the best with "best" defined as the patient with the highest cholesterol level.
b. Select the subset of patients that contains the patient with the highest cholesterol count with a probability of correct selection of .95.

6. A set of comparisons of interest on the chicken experiment in Exercise 2.3 was the comparison of serum T3 for each of the other states with that for the premolt stage.
a. Compute the 95% SCI comparisons of other stages with the premolt stage using the Dunnett method.
b. What are your conclusions?

7. Sections of tomato plant tissue were grown in tissue cultures with differing amounts and types of sugars in an experiment with five replications of four treatments in a completely randomized design. The tissue growth of each culture is given in the table below as mm $\times$ 10.

Control	*3% Glucose*	*3% Fructose*	*3% Sucrose*
45	25	28	31
39	28	31	37
40	30	24	35
45	29	28	33
42	33	27	34

a. Compute the 95% SCI comparisons of all treatments with the control treatment using the Dunnett method.
b. What are your conclusions?

8. The coefficients shown to you by a colleague for a set of contrasts among treatment means follow. He wants you to check them out.

Treatment	*A*	*B*	*C*	*D*	*E*
*C*1	1	3	−1	−1	−1
*C*2	1	−1	0	−1	1
*C*3	−1	1	−1	1	−1
*C*4	0	0	2	−1	−1

a. Does each of the proposed set of coefficients constitute a contrast? Justify your answer.
b. Are *C*1 and *C*2 orthogonal? Justify your answer.
c. Construct a contrast orthogonal to *C*4 that is different from the others already shown.

9. Use the Scheffé test at the .05 level of significance to test the null hypotheses about the contrasts in Exercise 3.1.

10. Use the Bonferroni t test at the .05 level of significance to test the null hypothesis about the contrasts in Exercise 3.1.

11. Use the data on serum T3 in Exercise 2.3.
a. Conduct all pairwise comparisons with the Tukey method at the .05 significance level.
b. Conduct all pairwise comparisons with the Least Significant Difference at the .05 significance level.
c. Conduct all pairwise comparisons with the SNK multiple-range test at the .05 significance level.
d. How did the results differ among the three tests?
e. Explain why the results differed.
f. Compute the 95% SCI for all pairwise comparisons with the Tukey method.
g. What additional information do you have after computing the 95% SCI?

12. The following is a description of an experiment on human work systems. In a human work system, such as a factory assembly line, workers are often required to move an object to a specific location with their hand.

The specific purpose of the study was to determine the accuracy with which individuals could reach to specific target locations on a horizontal plane (for example, the top of a table) with their field of view cut off from the targets.

Previous research led to the hypothesis that distal movements (movements away from the body) are more accurate than proximal movements (movements toward the body). It was also hypothesized that movements in the cardinal directions (straight ahead, straight toward the body, and lateral) were more accurate than movements in other, non-cardinal, directions.

Targets were set up on the circumference of a circle with a radius of 10 inches (Figure 3.4). The subject was seated so that a movement of the right hand to the 90° target position from the starting point was a distal (away from body) movement. A movement of the right hand to the 270° target was a proximal (toward body) movement. Movements to the 0° and 180° targets were lateral movements. Distal movements to the 45° and 135° targets were non-cardinal movements as were proximal movements to 225° and 315°. The 0°, 90°, 180°, and 270° targets were cardinal directions.

Sixteen right-handed male subjects were trained for the study. In the actual trial the investigator randomly called out the target angles to the subject. The subject marked his try at the target, which was blocked from his view. The distance from the subject's mark to the target was recorded for each target. The average distance to each target location was computed from the observations on the 16 subjects. A smaller average value represented greater accuracy.

Show in a table the coefficients required for a contrast among means for each of the following comparisons of accuracy between movement directions.

$C1$:	Distal versus proximal in general
$C2$:	Cardinal versus non-cardinal
$C3$:	Lateral versus distal
$C4$:	Lateral versus proximal
$C5$:	Non-cardinal distal versus non-cardinal proximal
$C6$:	Cardinal distal versus cardinal proximal

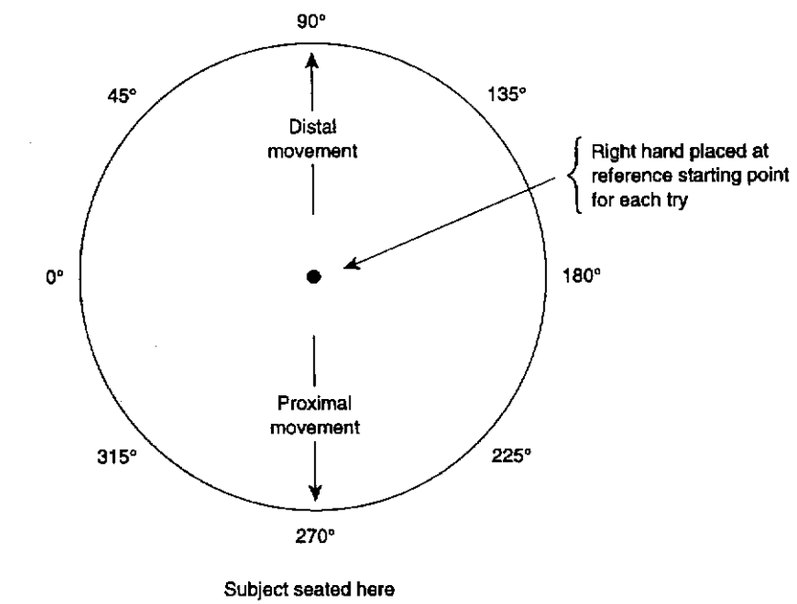

Figure 3.4 Targets set up on circle with radius of 10 inches

3A Appendix: Linear Functions of Random Variables

A linear function of the random variables $y_1, y_2, \cdots, y_n$ is defined as

$$c = \sum_{i=1}^{n} k_i y_i = k_1 y_1 + k_2 y_2 + \cdots + k_n y_n$$

If the expected value or mean of y_i is $E(y_i) = \mu_i$, then the expected value or mean of c is

$$\mu_c = E(c) = E\left(\sum_{i=1}^{n} k_i y_i\right) = \sum_{i=1}^{n} k_i E(y_i) = \sum_{i=1}^{n} k_i \mu_i$$

$$= k_1\mu_1 + k_2\mu_2 + \cdots + k_n\mu_n$$

The variance of a linear function $c = \Sigma k_i y_i$ is

$$\sigma_c^2 = \sum_{i=1}^{n} k_i^2 \sigma_i^2 + 2 \sum_{i<j}^{n} k_i k_j \sigma_{ij}$$

For example, if $c = y_1 - y_2$ with $k_1 = 1$ and $k_2 = -1$, then

$$\sigma_c^2 = k_1^2\sigma_1^2 + k_2^2\sigma_2^2 + 2k_1k_2\sigma_{12} = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

If the y_i's are independent, then $\sigma_{ij} = 0$ and

$$\sigma_c^2 = \sum_{i=1}^{n} k_i^2 \sigma_i^2$$

Therefore, if y_1 and y_2 are independent in $c = y_1 - y_2$, the variance of c is $\sigma_c^2 = \sigma_1^2 + \sigma_2^2$.

The Sample Mean

The mean of a sample of r independent observations from a normal distribution with a mean μ and variance σ^2 is a linear function

$$\overline{y} = \frac{1}{r} y_1 + \frac{1}{r} y_2 + \cdots + \frac{1}{r} y_r$$

where $k_i = \dfrac{1}{r}$. The expected value of the sample mean is

$$\mu_{\overline{y}} = E(\overline{y}) = \frac{1}{r}E(y_1) + \cdots + \frac{1}{r}E(y_r) = \frac{1}{r}(r\mu) = \mu$$

Since the variances of the observations are all the same, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$, then the variance of the mean is

$$\sigma_{\bar{y}}^2 = \frac{1}{r^2}(r\sigma^2) = \frac{\sigma^2}{r}$$

Linear Function of Sample Means

If t samples are independent and r_i is the number of observations in the ith sample, then a linear function of the sample means

$$c = k_1\bar{y}_1 + k_2\bar{y}_2 + \cdots + k_t\bar{y}_t$$

has a mean

$$\mu_c = E(c) = k_1E(\bar{y}_1) + k_2E(\bar{y}_2) + \cdots + k_tE(\bar{y}_t)$$

$$= k_1\mu_1 + k_2\mu_2 + \cdots + k_t\mu_t$$

and a variance

$$\sigma_c^2 = k_1^2\left(\frac{\sigma_1^2}{r_1}\right) + k_2^2\left(\frac{\sigma_2^2}{r_2}\right) + \cdots + k_t^2\left(\frac{\sigma_t^2}{r_t}\right)$$

If all sample variances are equal, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2$, then

$$\sigma_c^2 = \sigma^2\left(\frac{k_1^2}{r_1} + \frac{k_2^2}{r_2} + \cdots + \frac{k_t^2}{r_t}\right)$$

4 Diagnosing Agreement Between the Data and the Model

The analysis of variance can lead to erroneous inferences if certain assumptions regarding the data are not satisfied. Diagnostic methods for detecting faulty assumptions are discussed in Chapter 4 along with data transformations that can be used to address the problems. A generalization of the linear model for the analysis is suggested as an alternative to data transformations. Also, a graphical method is introduced to evaluate how well a model fits the data.

4.1 Valid Analysis Depends on Valid Assumptions

The validity of estimates and tests of hypotheses for analyses derived from the linear model rests on the merits of several key assumptions. The random experimental errors are assumed to be independent, be normally distributed with a mean of zero, and have a common variance (σ^2) for all treatment groups. Any disagreement between the data and one or more of these assumptions affects the estimates of the treatment means and tests of significance from the analysis of variance.

Summary discussions on the assumptions for the analysis of variance and effects of departures from the assumptions can be found in Eisenhart (1947) and Cochran (1947). Ito (1980) summarized research on the validity of analysis of variance test procedures under departures from assumptions.

4.2 The Effects of Departures from Assumptions

If experimental errors are positively correlated, Cochran (1947) showed that the actual precision of the treatment mean is less than the estimated precision. The