CHAPTER 9. MULTIPLE COMPARISONS AND TRENDS AMONG TREATMENT MEANS

The analysis of variance method is a useful and powerful tool to compare several treatment means. In comparing k treatments, the null hypothesis tested is that the k true means are all equal ($H_O : \mu_1 = \mu_2 = ... = \mu_k$). If a significant F test is found, one accepts the alternative hypothesis which merely states that they are not all equal. Further comparisons to determine which treatments are different can be carried out by using so-called multiple comparison procedures or by further partitioning of the treatment sum of squares to provide additional F tests to answer planned questions.

Before describing multiple comparison procedures, we will discuss the question of error rates. When comparing three or more treatments in an experiment, there are at least two kinds of type I error rates:

Comparison-wise type I error rate

= (number of type I errors)/(number of comparisons)

Experiment-wise type I error rate

= (number of experiments with one or more type I errors)/ (number of experiments)

If each experiment has only two treatments, these rates are identical.

Suppose an experimenter conducts 100 experiments with 5 treatments each. In each experiment there are $\binom{5}{2} = 10$ possible pairwise comparisons and in all experiments, 1000 comparisons. Assume there are no true differences among the 5 treatments but that in each experiment one mistake is made among the 10 comparisons, i.e., the rejection of the null hypothesis that there is no difference between 2 treatments. The comparison-wise error rate over all experiments is:

$$\frac{100 \text{ mistakes}}{1000 \text{ comparisons}} \bullet (100) = 10\%$$

The experiment-wise error rate is:

$$\frac{100 \text{ experiments with mistakes}}{100 \text{ experiments}} \bullet (100) = 100\%$$

Thus to preserve a low experiment-wise error rate, the comparison -wise error rate has to be kept extremely low. Conversely, to maintain a reasonable comparison-wise error rate, the experiment-wise error rate must be considerably larger.

The relative importance of controlling these two type I error ratios depends on the objectives of the study and the number of treatments involved. Different multiple comparison procedures have been developed based on different philosophies of controlling these two kinds

of errors. In selecting a procedure, there is no universal criterion that enables us to decide whether a comparison-wise or an experiment-wise error rate is more appropriate to be controlled.

In situations where incorrectly rejecting one comparison may jeopardize the entire experiment or the consequence of incorrectly rejecting one comparison is as serious as incorrectly rejecting a number of comparisons, then the control of experiment-wise error rate is most important. On the other hand, when one erroneous conclusion does not affect the remaining inferences in an experiment, the comparison-wise error rate is pertinent.

In most agricultural experiments, treatments can be planned to provide specific F tests for certain relationships among the treatment means. Multiple comparison procedures are useful in those experiments where there are no particular relationships among the treatment means.

Many pairwise comparison procedures are available and considerable controversy exists as to which procedure is most appropriate. We will present four commonly used procedures.

9.1 Pairwise comparison procedures

To illustrate the various procedures for pairwise comparisons, we will use the data given in Table 9-1 and 9-3, representing experiments with equal and unequal replications. The analysis of variance for these experiments are given in Tables 9-2 and 9-4.

Table 9-1. Results (mg shoot dry weight) of an experiment (CRD) to determine the effect of seed treatment by acids on the early growth of rice seedlings.

Treatmens	Replications					Total ($Y_{i.}$)	Mean ($\overline{Y}_{j.}$)
Control	4.23	4.38	4.1	3.99	4.25	20.95	4.19
HC1	3.85	3.78	3.91	3.94	3.86	19.34	3.87
Propionic	3.75	3.65	3.82	3.69	3.73	18.64	3.73
Butyric	3.66	3.67	3.62	3.54	3.71	18.2	3.64
Overall						$Y_{..}$=77.13	$\overline{Y}_{..}$= 3.86

Table 9.2. AOV of data in Table 9-1.

Source of Variation	df	Sum of Squares	Mean Squares	F
Total	19	1.0113		
Treatment	3	0.8738	0.2912	33.87
Exp. error	16	0.1376	0.0086	

Table 9-3. Heifer weight gains (lb/animal/day) as affected by three different feeding rations. CRD, unequal replications.

Treat-ment	Replications							Number/ treatment	Total (Yi.)	Mean ($\overline{Y}$ I.)
Control	1.21	1.19	1.17	1.23	1.29	1.14		6	7.23	1.20
Feed-A	1.34	1.41	1.38	1.29	1.36	1.42	1.37 1.32	8	10.89	1.36
Feed-B	1.45	1.45	1.51	1.39	1.44			5	7.24	1.45
Feed-C	1.31	1.32	1.28	1.35	1.41	1.27	1.37	7	9.31	1.33
Overall								26	Y.. = 34.67	
									$\overline{Y}$ .. = 1.33	

Table 9-4. AOV of data in Table 9-3.

Source of Variation	df	Sum of Squares	Mean Square	F
Total	25	0.2213		
Treatment	3	0.172	0.0573	25.57
Exp. error	22	0.0493	0.0022	

Fisher's Protected Least Significant Difference (PLSD)

Fisher (1935) described a procedure for pairwise comparisons called the least significant difference (LSD) test. This test is to be used only if the hypothesis that all means are equal is rejected by the overall F test. If the overall test is significant, a procedure analogous to ordinary Student's t test is used to test any pair of means. If the overall F ratio is not significant, no further tests are performed. When it is used, the two treatments will be declared different if the absolute difference between two sample means

(say $\overline{Y}_A$ and $\overline{Y}_B$) is greater than the LSD given by

$PLSD = t_{\alpha,df} S_{\bar{d}}$, where df is the degrees of freedom for experimental error.

$= t_{\alpha,df} \sqrt{2MSE / r}$, where r is the replication number fro each treatment.

$= t_{\alpha,df} \sqrt{MSE\ (1/r_A + 1/r_B}$,

if treatments are not equally replicated.

Note, when all the treatments are equally replicated, only one LSD value is required to test the 6 possible comparisons between the treatment means of Table 9-1. A different LSD must be calculated for each comparison involving different numbers of replications or 6 different LSD's for the uneven comparisons of Table 9-3.

One advantage of the PLSD procedure is its ease of application. Additionally, it is readily used to construct confidence intervals for mean differences, $\mu_A - \mu_B$. The 1- confidence limits are

$$\begin{matrix} L \\ U \end{matrix} = (\overline{Y}_A - \overline{Y}_B) \pm PLSD$$

Since the F value of Table 9-2 is highly significant we will use LSD for comparisons among treatment means of Table 9-1. Note that the α level selected for pairwise comparisons does not have to conform to the significance level of the overall F test. To compare procedures in the following examples, we use $\alpha = 0.01$.

From Table 9-1, MSE = 0.0086 with 16 df and

$t_{0.01,16} = 2.92$. Thus,

$$PLSD = 2.92\sqrt{2(0.0086)/5} = 0.171$$

If the absolute difference between any two treatment means is 0.171 or more, the treatments are said to be significantly different at the 1% level.

Identification of the pairs of treatments that are significantly different becomes increasingly difficult as the treatment number increases. A systematic procedure for comparison is to arrange the means in descending or ascending order as shown below.

Control	HC1	Propionic	Butyic
4.19	3.87 b	3.73 ab	3.64 a

First compare the largest with the smallest mean. If these two means are significantly different, then compare the next largest with the smallest. Repeat this process until a non-significant difference is found. Connect these two and any means in between with a common line or place a common lower case letter by each mean.

For the above example, we draw the following conclusions at the 1% level. All acids reduced shoot growth. The reduction was more severe with butyric acid than HC1. We do not have enough evidence to support a conclusion that propionic acid is different in its effect to either HC1 or butyric acid. (At the 5% level, however, the difference between HC1 and propionic acid is significant). The 99% confidence interval for the difference of any two means is

$$\begin{matrix} L \\ U \end{matrix} = \overline{d} \pm 0.17$$

For example, between control and HC1,

$$\begin{matrix} L \\ U \end{matrix} = 0.32 \pm 0.17 = \begin{matrix} 0.15 \\ 0.49 \end{matrix}$$

For the case of unequal replications (Table 9-3 and 9-4), the ranked means are,

Control	Feed-C	Feed-A	Feed-B
1-20 c	1.33 b	1.36 b	1.45 a

The 1% PLSD for comparing the control with Feed-B (the greatest mean difference) is,

$$PLSD = 2.82\sqrt{0.0022\ (1/6+1/5)} = 0.0801$$

The other required PLSD's are: B vs C = 0.0774, B vs A = 0.0754, A vs Control = 0.0714, A vs C = 0.0684, and C vs Control = 0.0736. Thus, at the 1% level, we conclude that Feeds A and C are equally effective, but all the other treatments are significantly different.

Duncan's Multiple Range Test (DMR)

Duncan (1955)used a different approach to compare means, called the multiple range test. To apply the method, instead of comparing the difference between any two means with a constant least significant difference, each pair of means is compared against a different critical value which depends on the ranks of these means in the ordered array.

The formula for calculating critical values is

$$DMR_p = Q_p \bullet S_{\bar{y}} = Q_p\sqrt{MSE / r}$$

Q_p is the tabular value from Appendix Table A-8 for a given α, df for experimental error, and the degree of separation of the means in the array.

If the means being compared are arranged in order of magnitude, adjacent means having a difference greater than DMR_2 are considered significantly different. The difference between the largest and smallest of three consecutive means is considered significant if that exceeds DMR_3, or, in general, the difference between the largest and the smallest of any p consecutive means is considered significant if it exceeds DMR_p. Always start the test with the extremes. Once two means are declared to be not significantly different, we can underline them and no further testing is done between means underscored by this line.

In the case where treatment replications are not equal, the following method can be used to approximate the overall replication number r.

$$r = \frac{1}{k-1}(\Sigma r_i - \frac{\Sigma r_i^2}{\Sigma r_i})$$

where r_i is the replication number for treatment i, and k is the number of treatments.

One disadvantage of the DMR is that it is not amenable to simultaneous interval estimation of the difference between means. Since DMP_p depends on the number of treatments

involved in defining the range between the largest and the smallest means, some pairs of means will have confidence intervals of different widths, even if all treatments are equally replicated.

To illustrate the use of DMR with equal replication for all treatments, we use data from Tables 9-1 and 9-2.

MSE = 0.0086, df = 16, r = 5, k = 4 and
$S_{\bar{y}} = \sqrt{0.0086 / 5} = 0.0415$

At the 1% level,

p:	2	3	4
Q_p:	4.13	4.31	4.43
DMR_p:	0.171	0.179	0.183

The results of mean comparisons are:

Control	HCL	Propionic	Butyric
4.19	3.87	3.73	3.64

In this situation, the difference between the control mean and butyric mean (4.19 - 3.64 = 0.55) is greater than $DMR_4 = 0.18$ which is the critical value for 4 means; both 4.19 = 3.73 = 0.46 and 3.87 - 3.64 = 0.23 are greater than $DMR_3 = 0.18$. In comparisons of the adjacent means, we find there is significant difference between control and HC1, but no other difference is significant. Thus, the same conclusions are drawn as with the protected LSD test.

For the data of Tables 9-3 and 9-4, where replications are unequal,

$$r = \frac{1}{4-1}\{(6+8+5+7) - \frac{(6^2+...+7^2)}{(6+...+7)}\}$$
$$= \frac{1}{3}(26 - \frac{174}{26}) = 6.4$$

and

$$S_{\bar{y}} = \sqrt{0.0022 / 6.4} = 0.0185$$

At the 1% level, the critical values are:

P:	2	3	4
Q_p:	3.96	4.13	4.24
DMR_p:	0.073	0.077	0.079

The results of mean comparisons are:

Control	Feed-C	Feed-A	Feed-B
1.20 c	1.33 b	1.36 b	1.45 a

Again the conclusions are the same as with the PLSD. That is, there is no significant difference between feed-A and feed-C, but all other differences are significant at the 1% level.

Scheffe's F test

If the overall F ratio is significant, Scheffe's (1953) method can be used to make comparisons between groups of means as well as all possible pairwise comparisons. Since this procedure allows for more kinds of comparisons, it is less sensitive in finding significant differences than other pairwise comparison procedures.

For pairwise comparison, Scheffe's F is

$$F_s = \left(\frac{\overline{X}_A - \overline{X}_B}{S_{\bar{d}}}\right)^2 / (k-1) = t^2 / (k-1)$$

where k = the number of treatments, and

$S^2_{\bar{d}}$ = 2MSE/r, for equal replications, or

= MSE $(1/r_A + 1/r_B)$, for unequal replications.

The required tabular F value for a significance test is based on (k-1) and the df for experimental error.

Another way to use Scheffe's test is to compare the mean difference with the following critical value called Scheffe's critical difference (SCD),

$$SCD = [(k-1) \bullet F_{\alpha,(k-1),df_{error}} \bullet S^2_{\bar{d}}]^{1/2}$$

That is if $\left|\overline{X}_A - \overline{X}_B\right| \geq SCD$ the difference will be declared significant at the given α level.

Scheffe's procedure is also readily used for interval estimation. The (1 -) confidence level for $(\mu_A - \mu_B)$, is

$$\begin{matrix} L \\ U \end{matrix} = (\overline{X}_A - \overline{X}_B) \pm SCD$$

For the rice seedling experiment in Table 9-1 and 9-2,

k-1 = 3, r = 5, $F_{1\%,3,16} = 5.29$,

$S_{\bar{d}} = \sqrt{2(0.0086)/5} = 0.00344$,

and

$$SCD = \sqrt{3(5.29)(0.00344)} = 0.233$$

The results of the pairwise mean comparisons are,

Control	HC1	Propionic	Butyric
4.19	3.87	3.73	3.64

Now the control is found to be significantly different from all of the acid treatments but there are no significant differences among the acid treatments. Note the SCD (0.233) is larger than the previously obtained PLSD (0.171) or DMR_4 (0.183) and thus gives more conservative results (the previously declared significance between HC1 and butyric treatments is no longer significant by Scheffe's test).

When the means to be compared are not based on equal replications, a different $S_{\bar{d}}$ is required for each comparison. Again, for the data of Table 9-3 and the MSE in Table 9-4, the F_S for control verses Feed-B is,

$$F_s = \frac{(1.20 - 1.45)^2}{0.0022(\frac{1}{6} + \frac{1}{5})} \bullet \frac{1}{(4-1)}$$

$$= \frac{0.0625}{0.00081} \bullet \frac{1}{3} = 25.72$$

The tabular $F_{1\%,3,22} = 4.82$, therefore, the difference is highly significant. The other calculated F_S are:

B vs C = 6.40, B vs A = 3.75, A vs Control = 13.28,
A vs C = 0.51, C vs Control = 8.24

The ranked means and the comparison results are,

Control	Feed-C	Feed-A	Feed-B
1.20 c	1.33 b	1.36 ab	1.45 a

Thus, Scheffe's test is less sensitive than PLSD and DMR in that the A versus B difference is no longer significant at the 1% level.

As mentioned, Scheffe's test is also used for arbitrary comparison among groups of means. For this purpose, we will use the general form for Scheffe's F test:

$$F_s = \frac{1}{k-1} \bullet \frac{(\Sigma C_i \overline{Y}_i)^2}{\Sigma(C_i^2 / r_i)} \bullet \frac{1}{MSE}$$

where F_S is compared against the tabular F for degrees of freedom k-1 and of error. The C_i's, are coefficients associated with the treatment means to be compared with the condition that $\Sigma Ci = 0$.

For example, comparing the control mean with the mean of the three feed treatments, the Ci's, r_i, Y_i and calculations are,

	Control		Feed-B	Feed-C	Total
C_i	3	-1	-1	-1	0
r_i	6	8	5	7	
$\overline{Y}_i$	1.2	1.36	1.45	1.33	
$C_i \overline{Y}_i$	3.6	-1.36	-1.45	-1.33	-0.54
C_i^2 / r_i	1.5	0.13	0.2	0.14	1.97

$$F_s = \frac{1}{4-1}\frac{(-0.54)^2}{1.97} \bullet \frac{1}{0.0022} = 22.43$$

The resulting F_S is greater than $F_{1\%,3,22} = 4.82$, therefore 1.20 is significantly smaller than 1.38, the average of Feeds A, B, and C.

Dunnett's Method

In certain experiments, it may only be desirable to compare a control with each of several other treatments, such as comparing a standard variety or chemical with several new ones. For this purpose Dunnett's method provides a test to control the experiment-wise error rate, concentrating on fewer comparisons. In this method a t value is calculated for each comparison. The tabular t value for determining statistical significance, however, is not the Student's t but a special t given in Appendix Table A-9(a and b). Let $\overline{Y}0$ represent the control mean with r_0 replications, then,

$$t = \left|\overline{Y}_i - \overline{Y}_0\right| / \sqrt{MSE(1/r_i + 1/r_0)}$$
$$= \left|\overline{Y}_i - \overline{Y}_0\right| / S_{\bar{d}}$$

If sample sizes are equal,

$$t = \left|\overline{Y}_i - \overline{Y}_0\right| / \sqrt{2MSE / r}$$

Another form of Dunnett's test is similar to LSD;

$$DSLD = t^*_{\alpha,k-1,df_{error}} \bullet \sqrt{2MSE / r}$$

where t* is the tabular value from Appendix Table A-9(a and b). This provides the least significant difference between a control and any other treatment.

The 1-α simultaneous differences at the 1% level between the control and the acid treatments in Table 9-1, we calculate

$$\text{DLSD} = 3.39\sqrt{2(0.0086)/5} = 0.1988$$

Note that the smallest difference between the control and any acid treatment is

Control - HC1 = 4.19 - 3.87 = 0.32

Since this difference is larger than DLSD, it is highly significant, and all other differences, being larger, are also highly significant.

The 99% simultaneous confidence intervals for all three differences are computed as

$$(\overline{Y}_0 - \overline{Y}_i) \pm \text{DLSD}$$

The limits of these differences are,

Control	-	butyric	=	0.32 ± 0.20
Control	-	HC1	=	0.46 ± 0.20
Control	-	propionic	=	0.55 ± 0.20

That is, we have 99% confidence that the three true differences fall simultaneously within the above ranges.

When treatments are not equally replicated, only $S_{\bar{d}}$ changes according to the sample size. For example, in Table 9-3, to compare control with feed-A,

$$t = |1.20 - 1.36| / \sqrt{0.002(1/6 + 1/8)} = 6.32$$

to compare with control with feed-B,

t = 0.265 / 0.0284 = 8.80 and

to compare control with feed-C,

t = 0.13 / 0.0261 = 4.98

From Appendix Table A-9b (two-sided test), we find $t^*_{1\%,3,22.}$ = 3.26, thus all the differences are highly significant.

Some remarks on pairwise comparison procedures

The multiple comparison procedures described above are applicable for a wide range of research situations. There are at least twenty other parametric procedures available for multiple comparisons in addition to many nonparametric and multivariate methods.

There is no consensus as to which one is the most appropriate procedure to recommend to all users. One main difficulty in comparing the procedures lies with the different kinds of type I error rates used, namely, experiment-wise versus comparison-wise. In fact, the difference in performance of any two procedures is likely to be due to the different type I error probabilities than to the techniques used. To a large extent, the choice of a procedure will be subjective and will hinge on a choice between a comparison-wise error rate (such as LSD and DMR) and an experiment-wise error rate (such as PLSD and Scheffe's test). Although not required, it may be desirable to apply the DMR procedure only after a significant F test, to reduce the comparison-wise error rate to an experiment-wise error rate. Scheffe's method provides a very general technique to test all possible comparisons among means. For just pairwise comparisons, Scheffe's method is not recommended as it is overly conservative. Dunnett's test should be used if the experimenter only wants to make comparisons between each of several treatments and a control.

In the above comparisons, we used the 1% α level for mean separation, since the significance observed for the AOV's in Tables 9-2 and 9-4 was 1%. It is not necessary, however, to separate means at the same level as observed in an AOV. In the above cases, mean separation at the 5% level would also be appropriate.

9.2 Orthogonal Single Degree of Freedom F Tests

In planning an experiment, the investigator always has specific questions to be answered. The treatments employed are designed to provide information and statistical tests to answer those questions. An experienced investigator will select treatments so that the treatment sum of squares can be partitioned to answer as many independent questions as there are degrees of freedom for treatments in the AOV. In these cases, pairwise comparison procedures may not be appropriate to answer such a set of questions.

Any test concerning a set of means can be defined as a linear combination of the means. There are two general kinds of linear combinations -- class comparisons and trend comparisons. Trend comparisons deal with the area of regression analysis and will be considered at the end of this chapter and in more detail in Chapter 10.

Class comparisons

Consider the rice seedling experiment of Table 9-1 and the AOV in Table 9-2. A set of questions the investigator may have designed the treatments to answer is,

1) Do acid treatments decrease seedling growth?

2) Are organic acids different from inorganic acids?

3) Is there a difference in the effects of the two organic acids?

To facilitate answering these questions, a table of coefficients is show (Table 9-5) to define the linear combinations among treatments.

Table 9-5. Orthogonal coefficients for partitioning the treatment sum of squares of Table 9-2 into three independent tests.

Treatments and Totals and Means

		Control	HC1	Propionic	Butyric
	Y_i .	20.95	19.34	18.64	18.2
Comparisons	$\overline{Y}_{i.}$	4.19	3.87	3.73	3.64
Control vs. acid		+3	-1	-1	-1
Inorganic vs. organic		0	-2	+1	+1
Between organics		0	0	+1	-1

These coefficients can be used to partition the sum of squares of treatments (SST) into three components each with 1 degree of freedom to provide for an F test for each of the comparisons of Table 9-5. The critical tabular F is based on 1 df for the numerator and the df for error in the denominator. These tests are called single degree of freedom F tests. Coefficients for the three comparisons are derived from the following simple rules.

1. In comparing the means of two groups, each containing the same number of treatments, assign +1 to the members of one group and -1 to the members of the other. Thus for line 3 in Table 9-5, we are comparing two means, and assign coefficients of 1 (of opposite sign) to each. The same procedure is extended to the case of more than one treatment in each group.

2. In comparing groups containing different numbers of treatments, assign to the first group coefficients equal to the number of treatments in the second group; to the second group, assign coefficients of opposite sign, equal to the number of treatments in the first group. Thus for the first comparison of Table 9-5, where the control mean is compared with the mean of the three acids, we assign a 3 to the control and a 1 to each of the three acids. Opposite signs are then assigned to the two groups. It is immaterial as to which group gets the positive or negative sign since the sum of squares of the comparison will be calculated and used to form a F-test. The coefficients for any comparison should be reduced to the smallest possible integers for each calculation. Thus, +4, +4, -2, -2, -2, -2. should be reduced to +2, +2, -1, -1, -1, -1.

3. At times, a comparison component may be an interaction of two other comparisons. The coefficients for this comparison are determined by multiplying the corresponding coefficients of the two comparisons. For example, in a fertilizer experiment with four treatments, 2 levels of N and 2 levels of P, the comparisons are shown below.

	N_OP_O	N_OP_1	N_1P_O	N_1P_1
Between N	-1	-1	1	1
Between P	-1	1	-1	1

Interaction (NxP)	1	-1	-1	1

The coefficients for the first two comparisons are derived by rule 1. The interaction coefficients are the result of multiplying the coefficients of the first two lines.

Note that the sum of the coefficients of each comparison is zero and that the sum of the cross products of any two comparisons is also zero. When these two conditions are met, the comparisons are said to be orthogonal. This implies that the conclusion drawn for one comparison is independent of (not influenced by) the others.

The computation of the sum of squares for a single degree of freedom F test is

$$SS_1 = (\Sigma C_i Y_i.)^2 / r(\Sigma C_i^2) \quad \text{or}$$

$$= r(\Sigma C_i \overline{Y}_i.)^2 / \Sigma C_i^2$$

Referring to Table 9-5, we compute sum of squares for the comparisons and summarize them in Table 9-6.

SS_1 (control vs. acid)

$$= [3(30.95) - 18.20 - 18.64 - 19.34]^2/[5(12)]$$

$$= 0.7415$$

SS_1 (Inorg. vs. org.)

$$= [18.20 + 18.64 -2(19.34)]^2/[6(5)]$$

$$= 0.1129$$

SS_1 (between org.)

$$= (-18.20 + 18.64)^2/ [2(5)]$$

$$= 0.0194$$

Table 9-6. Orthogonal partitioning of treatments of Table 9-2.

Source of Variation	df	SS	MS	F
Total	19	1.0113		

Treatment	3	0.8738	0.2912	33.87
Control vs. acid	1	0.7415	0.7415	86.22
Inorg. vs. Org.	1	0.1129	0.1129	13.13
Between Org.	1	0.0194	0.0194	2.26
Error	16	0.1376	0.0086	

From the above analysis we conclude that in this experiment all three acids significantly reduce seedling growth ($p<0.01$), that organic acids cause more reduction than the inorganic acid ($p<0.01$) and that the difference between the organic acids is not significant ($p<0.05$).

Note that the single degree of freedom sum of squares add up to the sum of squares of treatments. This will always happen when the individual comparisons are orthogonal. The maximum number of orthogonal comparisons equals the degrees of freedom for the treatment sum of squares. When comparisons are not orthogonal, the sum of squares for one comparison may contain (or be contained by) part of the sum of squares of another comparison. Therefore, the conclusion from one test may be influenced by another test and the sum of squares of those individual comparisons will not add up to the sum of squares for treatments.

Trend comparisons

Experiments are often designed to study the effect of increasing levels of a factor, e.g., increments of a fertilizer, planting dates, doses of a chemical, concentrations of a feed additive, etc. In these situations, the experimenter is interested in the dose response relationship. The statistical analysis should evaluate the trend of the response and not be concerned with pairwise comparisons.

To illustrate the procedure to evaluate a trend response, we will use the data of Table 8-4 and the AOV of Table -6, and will partition the sum of squares for treatments into a linear and a residual component. By computing a single degree of freedom F test for the linear component, we will be able to determine whether the % sucrose declines linearly as the N-fertilizer increases.

The AOV for the experiment is repeated in Table 9-7 and includes the parititoned sum of squares for treatments.

Table 9-7. AOV of % sucrose of sugar beets treated with N-fertilizer.

Source of Variation	df	SS	MS	F
Total	59	62.25		
Plots	29	55.71		
Blocks	4	9.53	2.38	4.25

Treatments	5	34.94	6.99	12.48
Linear	1	33.18	33.18	75.41
Residual	4	1.76	0.44	0.79
Exp. error	20	11.24	0.56	2.43
Sampling error	30	6.94	0.23	

Treatment totals and the coefficients for the linear comparison are,

	N_O	N_{50}	N_{100}	N_{150}	N_{200}	N_{250}
$Y_{i..}$	161.6	157.4	152.9	152.9	143.6	139.4
C_i	-5	-3	-1	+1	+3	+5

The C_i's are found in Appendix Table A-10. Note that the coefficients provided in the table are for equally spaced treatment levels. The evaluation of a trend analysis is considerably simplified when treatment levels are equally spaced either in arithmetic or logarithmic scales.

The sum of squares for N-linear is calculated from

$$SSL = (\Sigma C_i Y_{i.})^2 / (rs\Sigma C_i^2)$$
$$= [-5(161.6) + ... + 5(139.4)]^2/[5(2)(70)]$$
$$= 33.18$$

The sum of squares for N-residual is calculated from

$$SSR = SST - SSL = 34.94 - 33.18 = 1.76$$

As there are 5 df for treatments, the SST can be partitioned into 5 single degree of freedom trend comparisons, linear, quadratic, cubic, etc. Since the residual sum of squares (1.76) is so small that even if it were entirely associated with one single degree of freedom component it still would not be significant, no further partitioning is necessary. Note that the linear trend accounts for 97% (33.18/34.23) of the variability due to N fertilizer. Therefore it is this tendency of % sucrose to decrease with increasing amounts of N fertilizer that is important and not pairwise comparisons. The highly significant F value for a linear trend implies that any increase in fertilizer N will cause a decease in % sucrose. This means that all 6 nitrogen treatments are significantly different from each other even though some observed means may be the same.

The next step is to estimate the rate of decrease in % sucrose per unit increase of N application. This is a problem in linear regression and the discussion will be deferred until Chapter 10.

In the next example, we will illustrate how the trend comparison procedure is applied to a factorial experiment. Recall the experiment of Table 8-9 and 8-10. In this case, we have 2 levels of irrigation and 5 levels of N-fertilizer. The planned questions for this study may be:

1. Is there a difference between irrigation levels?

2. How does the yield respond to increasing levels of N? (linear, quadratic, cubic or quartic?)

3. Does the irrigation level affect the response curves? (I x N-linear, I x N-quadratic, I x N-cubic, or I x N-quartic?)

A table of the orthogonal coefficients to answer the above questions is given in Table 9-8. The first question is a class comparison, one irrigation level versus another, and coefficients are determined according to the rules presented. Coefficients for response trends for N-levels are found in Appendix Table A-10.

Table 9-8. Orthogonal coefficients for partitioning the treatment sum of squares of Table 8-8.

	I_1					I_2				
	N_0	N_{80}	N_{160}	N_{240}	N_{320}	N_0	N_{80}	N_{160}	N_{240}	N_{320}
$Y_i.$	70.8	98.6	117.4	116.7	107.2	78.9	114.5	135.2	146.4	142.9
Comparison $\overline{Y}_{i.}$	35.4	49.3	58.7	58.4	53.6	39.5	57.3	67.7	73.2	71.5
Irrigation	1	1	1	1	1	-1	-1	-1	-1	-1
N-linear	-2	-1	0	1	2	-2	-1	0	1	2
N-quadratic	2	-1	-2	-1	2	2	-1	-2	-1	2
N-cubic	-1	2	0	-2	1	-1	2	0	-2	1
N-quartic	1	-4	6	-4	1	1	-4	6	-4	1
IxN-linear	-2	-1	0	1	2	2	1	0	-1	-2
IxN-quadratic	2	-1	-2	-1	2	-2	1	2	1	-2
IxN-cubic	-1	2	0	-2	1	1	-2	0	2	-1
IxN-quartic	1	-4	6	-4	1	-1	4	-6	4	1

The coefficients for any one interaction term are determined by multiplying corresponding coefficients for the two components involved as was done for class comparisons.

Again, to compute a single degree of freedom sum of squares we use the formula,

$$SS_1 = (\Sigma C_i Y_i.)^2 / (r \Sigma C_i^2)$$

and enter the results in Table 9-9. For example,

$$SS_I = (70.8 + 98.6 + ... - 142.9)^2/[2(10)] = 574.59$$

As another example,

$$SS_{I \times N\text{-linear}} = [-2(70.8) + ... -2(142.9)]^2/[2(20)] = 119.03$$

Table 9-9. AOV of Table 8-8 with the treatment sum of squares partitioned into single degree of freedom comparisons.

Source of variation	df	Sum of squares	Mean square	F
Total	19	2937.66		
Block	1	1.25	1.25	<1
Treatment	9	2861.08	317.9	37.98
Irrigation	1	574.59	574.59	68.64
(Nitrogen	4	2163.12	540.78	64.61)
linear	1	1572.52	1572.52	187.87
quadratic	1	590.2	590.2	7.07
cubic	1	0	0	<1
quadratic	1	0.40	0.40	<1
(IxN)	4	123.37	30.84	3.68
IxN-linear	1	119.03	119.03	14.22
IxN-quadratic	1	0.73	0.73	<1
IxN-cubic	1	0.00	0.00	<1
IxN-quartic	1	3.61	3.61	<1
Exp. error	9	75.33	8.37	

Note that the nine single degree of freedom sum of squares add up to the total treatment sum of squares. Each of the nine F tests provides an answer to one of the planned questions. As shown in Table 8-10, there were differences due to irrigation and nitrogen rates and response to nitrogen depended upon irrigation level. We can now examine the shape of the response to nitrogen and the nature of the interaction. For five nitrogen levels, it is possible to fit up to a quartic polynomial response curve. The significance of the coefficients in the polynomial model are tested by the F value calculated for each. In this case, only two coefficients are found to be significant, linear and quadratic, suggesting that a quadratic response model is adequate to describe the responses over the N-levels. The breakdown of the interaction sum of squares indicates that the responses to nitrogen for the two irrigation levels are only significantly different with respect to the linear coefficients. Therefore two response curves should be constructed, as in Figure 8-1, one for each irrigation level. The difference between these two models is primarily due to their linear components.

The analysis of Table 9-9 reveals far more information concerning the treatment effects than can be obtained by pairwise comparisons. The methods for fitting the appropriate regression equations to the response shown in Figure 8-1 will be discussed in Chapter 10.

Some remarks on treatment levels for trend analysis

The selection of dose levels for a material depends on the objectives the experimenter wishes to accomplish. If it is known that a certain response is linear over a given dose range and one is only interested in the rate of change, two doses will suffice a given dose range and one is only interested in the rat of change, two doses will suffice -- one low and one high. However, with only 2 doses there is no information available to verify the assumption of linearity. It is good practice to use one extra level so that the deviation from linearity can be estimated and tested. Similarly, if a quadratic response is expected, a minimum of 4 dose levels are required to test whether or not a quadratic would be appropriate.

The variability in agricultural data is generally greater than for physical and chemical laboratory studies, as the experimental units are subject to less controllable environmental influences. Also, responses vary from year to year and location to location. These variations cause difficulty in analyzing and interpreting combined experiments that are conducted over a few years or for a few locations. Furthermore, true response models are rarely known. For these reasons agricultural experiments require several dose levels, usually 4 to 6 levels to characterize a dose-response curve. It is usually desirable to have the doses equally spaced, arranging the levels in an arithmetic or logarithmic series. Experiments with equally spaced doses have the advantages of easy analysis and more power in determining the true response curve.

SUMMARY

Pairwise Comparison Procedures

The formula for calculating critical values of different procedures are shown below. Two treatments are declared different if the absolute difference between the two means is greater than the critical value according to the test.

1. Protected least significant difference test

$$PLSD = t_{\alpha/2,df}\sqrt{2MSE / r}$$ where r is the replication number for each treatment

$$= t_{\alpha/2,df}\sqrt{MSE(1/r_A + 1/r_B)}$$ if treatments are not equally replicated

t is the tabular value from Appendix Table A-6.

2. Duncan's Multiple Range Test

$$DMR_p = Q_p \bullet S_{\bar{r}} = Q_p\sqrt{MSE / r}$$ where Q_p is the tabular value from Appendix Table A-8.

Let $r = (\Sigma r_i - \Sigma r_i^2/\Sigma r_i)/(k-1)$

if treatment replications are not equal.

3. Scheffe's F test

$$SCD = \sqrt{(k-1) \bullet F_{\alpha,k-1,df_{error}} \bullet S_{\bar{d}}^2}$$

F is the tabular value from Appendix Table A-7.

4. Dunnett's method

$$DLSD = t^*_{\alpha,k-1,df_{error}} \bullet \sqrt{2MSE / r}$$

where t^* is the tabular value from Appendix Table A-9 (a and b).

Orthogonal Single Degree of Freedom F Tests

1. Class comparisons are comparisons between groups of means. The formula for calculating the single degree of freedom sum of squares is:

$$SS_1 = (\Sigma C_i Y_i.)^2 / r(\Sigma C_i^2) = r(\Sigma C_i \bar{Y}_i.)^2 / \Sigma C_i^2$$

where C_i's are coefficients to be determined according to the rules stated previously.

2. Trend comparisons are applied to test hypotheses regarding the patterns of the response over treatment levels. The formula for calculating the sum of squares for each trend analysis is the same as for the class comparison. However, the coefficients are determined from Appendix Table A-10.

3. There are as many single degree of freedom tests that can be made as there are degrees of freedom for treatments in the experiment.

4. Orthogonal tests are preferable rather than pairwise comparisons whenever applicable.

EXERCISES

1. Give an example of an experiment in your major field, where a multiple comparison procedure can be appropriately used to separate means. Indicate which procedure you will use and reasons why.

2. Describe an experiment to compare means where there is a difference in results when using a multiple comparison procedure controlling the experiment-wise error rate and a multiple comparison procedure controlling the comparison-wise error rate. Which is more appropriate?

3. Uterine weights (mg) of four mice for each control and six different solutions were measured to assay the estrogenic activity. The following information were obtained:

Treatment:	Control	1	2	3	4	5	6
Sample size:	8	4	4	4	4	4	4
$\overline{Y}_i$:	96	94	75	69	87	80	66

MSE = 78.5 with 25 degrees of freedom

Determine which means are significantly different by LSD, DMR and Scheffe's methods at the 1% significance level. Also use Dunnett's method to find out which treatment mean is significantly different from the control mean at the 1% level.

4. Refer to problem 6 of Chapter 7. Assume that there are some varietal differences. Use PLSD and DMR to identify which varieties are different. Use Scheffe's method to test whether the overall yield effect of varieties A, C, and E is different from varieties B and D at the 5% level. Assume variety A is the control and use Dunnett's test to compare A with the other varieties.

5. Refer to problem 11 of Chapter 7. If the multiple comparison between means is required, which procedure would you recommend? Perform the comparison according to your recommended procedure.

6. Refer to problem 12 of Chapter 7. If the multiple comparison between means is required, which procedure would you recommend? Perform the comparison according to your recommended procedure.

7. Refer to problem 1 of Chapter 8. If one wants to separate mean scores of products, which one of the multiple comparison procedures would you suggest? Use the recommended procedure to perform the comparison.

8. Refer to problem 14 of Chapter 8. Identify the states that are significantly different from one another.

9. Refer to problem 15 of Chapter 8. Identify the machines that are different from one another.

10. Refer to problem 14 of Chapter 7. How will you identify which treatments are different from one another? Use that method to perform the analysis.

11. Refer to problem 3 of Chapter 8. Partition the nitrogen sum of squares to test hypotheses related to response trends.

12. Refer to problem 5 of Chapter 8. Perform statistical analyses to answer the question, are there specific patterns (linear, quadratic, etc.) of body weight gains related to litter number?

13. Refer to problem 17 of Chapter 8. Suggest a set of orthogonal comparisons among treatment means that may be interesting to the experimenter. Test the significance of those comparisons.

14. Refer to problem 18 of Chapter 8. Perform single degree of freedom F tests to determine whether there is a linear and/or a quadratic response curve of counts over time after blending.

15. Refer to problem 20 of Chapter 8. Is the body weight gain linearly or quadratically related to the percent corn in the ration? Are the response curves parallel between the 2 breeds?