

## Topic 5. Mean separation: Multiple comparisons [S&T Ch.8 except 8.3]

### 5. 1. Basic concepts

In the analysis of variance, the null hypothesis that is tested is always that all means are equal. If the F statistic is not significant then this hypothesis is not rejected and there is nothing more to do, except possibly try to make the experiment itself more sensitive. If the null hypothesis *is* rejected then at least one mean is significantly different from at least one other one. The ANOVA gives no indication of which means are significantly different. If there are only two treatments then there is no problem, but if there are more than two treatments then the problem remains of trying to determine which means are significantly different. This is the process of mean separation.

Mean separation takes two forms:

- Planned F test (Orthogonal F test, Topic 4).
- Effects suggested by the data (multiple comparison tests, Topic 5)

Of these, orthogonal F tests are preferred since they can provide for more precise separation than multiple comparison tests. In order to use them, however, some statisticians feel that there must be some predetermined grouping of means to be compared (for this reason they are called *planned* F tests). In most agricultural experiments, treatments can be planned to provide specific F tests for certain relationships among the treatment means. If no such grouping is known, then it can be generated by a preliminary experiment using multiple comparison tests. Multiple comparisons involve more than one comparison among three or more means and are particularly useful in those experiments where there are no particular relationships among the treatment means.

### 5. 2. Error rates

The selection of the most appropriate mean separation test is heavily influenced by the **error rate**. Recall that in a Type I error, the null hypothesis  $H_0$  is incorrectly rejected when it is actually true. The **Type I error rate** is the fraction of times a Type I error is made. In a single comparison this is the value  $\alpha$ . When comparing three or more treatment means there are at least two kinds of type I error:

#### Comparison-wise type I error rate CER

This is the number of type I errors divided by the total number of comparisons

#### Experiment-wise type I error rate EER

This is the number of experiments in which **at least** one type I error occurs divided by the total number of experiments

Suppose the experimenter conducts 100 experiments with 5 treatments each. Then it turns out that in each experiment there are 10 possible pairwise comparisons ( $(t-1)/2$ ) and therefore there are a total of 1000 possible comparisons. Suppose that there are

no true differences among the treatments (i.e.  $H_0$  is true). Suppose that in each of the 100 experiments one Type I error is made. Then the CER over all experiments is:

$$\text{CER} = (100 \text{ mistakes}) / (1000 \text{ comparisons}) = 0.1 \text{ or } 10\%.$$

The EER is

$$\text{EER} = (100 \text{ experiments with mistakes}) / (100 \text{ experiments}) = 1 \text{ or } 100\%.$$

The EER is the probability of a Type I error for the experiment. As the number of means increases, the chance of making at least one Type I error approaches 1. 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.

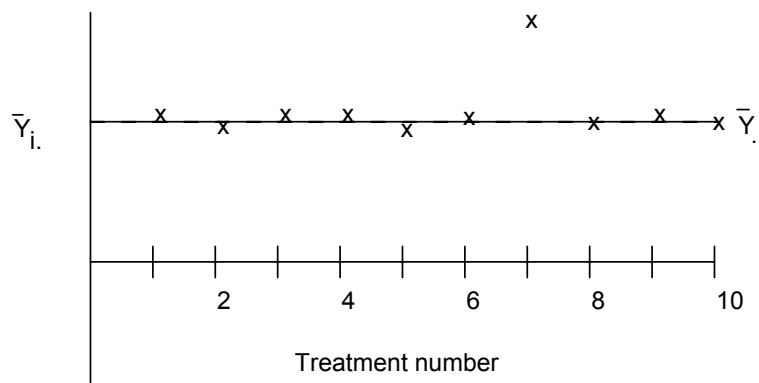
The relative importance of controlling these two Type I error ratios depends on the objectives of the study and the number of treatments involved. Different multiple comparison procedures have been developed based on different philosophies of controlling these two kinds of errors. In selecting a procedure, 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.

The experiment-wise error rate is larger than the comparison-wise error rate. It is difficult to compute the experiment-wise error rate because Type I errors are not independent, but it is possible to compute an upper bound by assuming that the probability of a Type I error in any comparison is  $\alpha$  and is independent of the other comparison probabilities. In this case:

$$\text{Upper bound EER} = 1 - (1 - \alpha)^p \text{ where } p = N^o \text{ of pairs to be compared } (p = t(t-1)/2).$$

If there are 10 treatments and  $\alpha = 0.05$  then the upper bound is 0.9.



The situation is more complicated than this, however. Suppose there are 10 treatments and one shows a significant effect while the other 9 are approximately equal. A simple ANOVA will probably reject  $H_0$ , so the experimenter will want to determine what the significant differences are. There is a probability of making a Type I error between each of the 9 similar treatments, and an upper bound on this probability is computed by setting  $t = 9$  in the above formula. It is 0.84, which is therefore the EER. That is, the experimenter will incorrectly conclude that some pair of similar effects are actually different 84% of the time. Thus, the experimenter must distinguish between the EER under the complete null hypothesis, in which all population means are equal, and the EER under a partial null hypothesis, in which some means are equal but some differ. Because of this fact SAS subdivides the error rates into:

**CER** = comparison-wise error rate

**EERC** = experimentwise error rate under complete null hypothesis (the standard EER)

**EERP** = experimentwise error rate under a partial null hypothesis.

**MEER** = maximum experimentwise error rate under any complete or partial null hypothesis.

### 5.3. Multiple comparison tests

Statistical methods for making two or more inferences while controlling the probability of making at least one Type I error are called *simultaneous inference methods*. This material is based primarily on ST&D chapter 8 and on the SAS/STAT manual (GLM Procedure). The basic techniques of multiple comparisons divide themselves into two groups: those which can provide confidence intervals and tests of hypothesis and those which are essentially only tests of hypothesis.

To illustrate the various procedures, we will use the data given in Table 4-1 and 5-1, representing experiments with equal and unequal replications. The analysis of variance for these experiments are given in Tables 4-2 and 5-2.

Table 5.1. Weight gains (lb/animal/day) as affected by three different feeding rations. CRD, with unequal replications.

Treatment								N	Total	Mean	
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	34.67	1.33	

Table 5-2. ANOVA of data in Table 5-1.

Source of Variation	df	Sum of Squares	Mean Squares	F
Total	25	0.2202		
Treatment	3	0.1709	0.05696	25.41
Exp. error	22	0.0493	0.00224	

### 5. 3. 1. Fixed-range tests

These tests provide one range for testing all differences in balanced designs and can provide confidence intervals. Many fixed-range procedures are available and considerable controversy exists as to which procedure is most appropriate. We will present four commonly used procedures starting from the less conservative to the more conservative: LSD, Dunnett, Tukey and Scheffe. There are other pairwise tests discussed in the SAS manual.

#### 5. 3. 1. 1. The repeated t and least significant difference: LSD

One of the oldest multiple pairwise comparison tests is the least significant difference (LSD) test. It is also one of the simplest and one of the most widely misused. The LSD is based on the t-test (ST&D p101). Recall that the formula for the t statistic with r degrees of freedom is

$$t = (\bar{Y} - \mu) / s_{\bar{Y}} \quad \text{With } s_{\bar{Y}} = \sqrt{s^2 / n}$$

The LSD test declares the difference between means  $\bar{Y}_i$  and  $\bar{Y}_{i'}$  of treatments i and i' to be significant when

$$|\bar{Y}_i - \bar{Y}_{i'}| > \text{LSD where}$$

$$\text{LSD} = t_{\alpha/2, \text{MSE df}} \sqrt{\text{MSE} \left( \frac{1}{r_1} + \frac{1}{r_2} \right)} \quad \text{for unequal r (SAS refers to this as the repeated t test)}$$

$$\text{LSD} = t_{\alpha/2, \text{MSE df}} \sqrt{\text{MSE} \frac{2}{r}} \quad \text{for equal r (SAS refers to this as the LSD test)}$$

Where MSE = pooled  $s^2$  and can be calculated by PROC ANOVA or PROC GLM.

The above statistic is called the studentized range statistic. The quantity under the square root is called the standard error of the difference or SED. For example, here are the calculations for Table 4.1. Note that the 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  $\alpha = 0.05$ . From Table 4-1, MSE = 0.0086 with 16 df and

$$\text{LSD}_{0.025} = 2.12 \sqrt{0.0086 \frac{2}{5}} = 0.1243$$

If the absolute difference between any two treatment means is 0.1243 or more, the treatments are said to be significantly different at the 5% 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.

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. Identify these two and any means in between with a common lower case letter by each mean.

Table 5.5

Treatment	Mean	LSD
Control	4.19	a
HCl	3.87	b
Propionic	3.73	c
Butyric	3.64	c

For the above example, we draw the following conclusions at the 5% level. All acids reduced shoot growth. The reduction was more severe with butyric and propionic acid than HCl. We do not have enough evidence to support a conclusion that propionic acid is different in its effect to butyric acid.

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 4-1. Different LSD must be calculated for each comparison involving different numbers of replications or 6 different LSD's for the uneven comparisons of Table 5.1. For this case:

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

The 5% LSD for comparing the control with Feed-B is,

$$\text{LSD}_{0.025} = 2.074 \sqrt{0.00224 \left( \frac{1}{6} + \frac{1}{5} \right)} = 0.0595$$

The other required LSD's are: B vs. C = 0.0575, B vs. A = 0.056, A vs. Control = 0.0531, A vs. C = 0.0509, and C vs. Control = 0.0546. Thus, at the 5% level, we conclude that Feeds A and C are equally effective, but all the other treatments are significantly different.

One advantage of the LSD procedure is its ease of application. Additionally, it is readily used to construct confidence intervals for mean differences,  $\mu_A - \mu_B$ . The 1- $\alpha$  confidence limits are  $= \bar{Y}_A - \bar{Y}_B \pm \text{LSD}$

The LSD test is much safer when the means to be compared are selected *in advance* of the experiment, although hardly anyone ever does this. It is primarily intended for use when there is no predetermined structure to the treatments (e.g. in variety trials). If a large number of means are to be compared and the ones compared are selected after the analysis of variance and the comparison focuses on means with most different values, then the actual error rate will be much higher than predicted.

The LSD test is the only test for which the error rate equals **the comparison wise error rate**. This is often regarded as too liberal (i.e. too ready to reject the null hypothesis). It has been suggested that the EEER can be held to the  $\alpha$  level by performing the overall ANOVA test at the  $\alpha$  level and making further comparisons only if the F test is significant (**Fisher's Protected LSD test**). However, it was then demonstrated that this assertion is false if there are more than three means. A preliminary F test controls the EERC but not the EERP

### 5. 3. 1. 2. 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. Dunnett's test holds the maximum experimentwise error rate under any complete or partial null hypothesis (MEER) to a level not exceeding the stated  $\alpha$ . 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 ST&D p 624-625. Let  $\bar{Y}_o$  represent the control mean with  $r_o$  replications, then:

$$DLSD = t^*_{\alpha/2, MSE \text{ df}} (\text{Dunnett}) \sqrt{MSE \frac{2}{r}}, \text{ for equal } r$$

$$DLSD = t^*_{\alpha/2, MSE \text{ df}} (\text{Dunnett}) \sqrt{MSE \left( \frac{1}{r_1} + \frac{1}{r_2} \right)} \text{ for unequal } r$$

From Table 4-1,  $MSE = 0.0086$  with 16 df and  $P=3$ ,  $t^*_{\alpha/2}=2.59$  ST&D p625

$$DLSD_{0.025} = 2.59 \sqrt{0.0086 \frac{2}{5}} = 0.152 \text{ (Note that } DLSD = 0.152 > LSD = 0.124)$$

This provides the least significant difference between a control and any other treatment. 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 significant, and all other differences, being larger, are also significant.

The 95% simultaneous confidence intervals for all three differences are computed as  $\bar{Y}_o - \bar{Y}_i \pm DLSD$ . The limits of these differences are,

Control	-	butyric	=	$0.32 \pm 0.15$
Control	-	HC1	=	$0.46 \pm 0.15$
Control	-	propionic	=	$0.55 \pm 0.15$

That is, we have 95% confidence that the three true differences fall **simultaneously** within the above ranges.

When treatments are not equally replicated, as in Table 5-1, to compare control with feed-C,  $t^*_{0.025, 22, P=3} = 2.517$  (from SAS) (Table A9-b = 2.54 or 2.51 for 20 and 24 df.)

$$DLSD = 2.517 \sqrt{0.00224 \left( \frac{1}{6} + \frac{1}{7} \right)} = 0.06627$$

Since  $\bar{Y}_o - \bar{Y}_c = 0.125$  is larger than 0.06627, it is significant. The other differences are also significant.

### 5. 3. 1. 3. Tukey's w procedure

Tukey's test was designed specifically for **pairwise comparisons**. The test, sometimes called "honestly significant difference test", controls the MEER when the sample sizes are equal. It uses a statistic similar to the LSD but with a number  $q_{\alpha,(p, \text{MSE df})}$  that is obtained from Table A8. Tukey critical value is larger than that of Dunnett because the Tukey family of contrasts is larger (all pairs of means).

$$w = q_{\alpha,(p, \text{MSE df})} \sqrt{\frac{MSE}{r}} \quad \text{for equal } r$$

We are not multiplying MSE x 2 because the Table A8 already includes the multiplication of its values by  $\sqrt{2}$ . For example for  $p=2$ , d.f.= infinite (similar to Normal distribution), and  $\alpha= 5\%$ , the critical value is 2.77, which is equal to  $1.96 * \sqrt{2}$

$$w = q_{\alpha,(p, \text{MSE df})} \sqrt{MSE \left( \frac{1}{r_1} + \frac{1}{r_2} \right) / 2} \quad \text{for unequal } r \text{ (SAS manual)}$$

For Table 4.1:  $q_{0.05,(4, 16)} = 4.05$

$$w = 4.05 \sqrt{\frac{0.0086}{5}} = 0.168 \quad \text{(Note that } w = 0.168 > \text{DLSD} = 0.152)$$

Table 5.6

Treatment	Mean	w
Control	4.19	a
HCl	3.87	b
Propionic	3.73	b c
Butyric	3.64	c

This test does not detect significant differences between HCl and Propionic (compare with Table 5.5).

For unequal r, as in Table 5.3, the contrast between control with feed-C,

$$q_{0.05,(4, 22)} = 3.93 \quad w = 3.93 \sqrt{0.00224 \left( \frac{1}{6} + \frac{1}{7} \right) / 2} = 0.0731$$

Since  $\bar{Y}_o - \bar{Y}_c = 0.125$  is larger than 0.0731, it is significant. As in the LSD the only pairwise comparison that is not significant is between Feed-C ( $\bar{Y}=1.33$ ) and Feed-A ( $\bar{Y}=1.36$ ).

### 5.3.1.4. Scheffe's F test

Scheffe's test is compatible with the overall ANOVA  $F$  test in that it never declares a contrast significant if the overall  $F$  test is nonsignificant. Scheffe's test controls the MEER for **ANY** set of contrasts including 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 comparisons with equal  $r$ , the Scheffe's critical difference SCD has a similar structure as that described for previous tests.

Recall that  $LSD = t_{\alpha/2, MSE\ df} \sqrt{MSE \frac{2}{r}}$ . Scheffe's SCD is:

$$SCD = \sqrt{df_{TR} F_{\alpha}(df_{TR}, df_{MSE})} \sqrt{MSE \frac{2}{r}} \text{ for equal } r$$

$$SCD = \sqrt{df_{TR} F_{\alpha}(df_{TR}, df_{MSE})} \sqrt{MSE \left( \frac{1}{r_1} + \frac{1}{r_2} \right)} \text{ for unequal } r$$

From Table 4-1,  $MSE = 0.0086$  with  $df_{TR} = 3$ ,  $df_{MSE} = 16$ , and  $r = 5$

$$SCD_{0.05} = \sqrt{3 * 3.24} \sqrt{0.0086 \frac{2}{5}} = 0.183 \text{ (Note that } SCD = 0.183 > w = 0.168)$$

If the mean difference  $\geq SCD$ , the difference will be declared significant at the given  $\alpha$  level. We can calculate Table 5.7:

Treatment	Mean	$F_s$
Control	4.19	a
HCl	3.87	b
Propionic	3.73	b c
Butyric	3.64	c

When the means to be compared are not based on equal replications, a different SCD is required for each comparison. For Table 5.1, the contrast Control vs. Feed-C,

$$SCD_{0.05, (3, 22)} = \sqrt{3 * 3.05} \sqrt{0.00224 \left( \frac{1}{6} + \frac{1}{7} \right)} = 0.0796$$

Since  $\bar{Y}_o - \bar{Y}_c = 0.125$  is larger than 0.0796, it is significant.

Scheffe's procedure is also readily used for interval estimation. The  $1 - \alpha$  confidence limits are  $= \bar{Y}_A - \bar{Y}_B \pm SCD$ . The resulting intervals are **simultaneous** in that the probability is at least  $1 - \alpha$  that all of them are simultaneously true.

The most important use of Scheffe's test is for arbitrary **comparisons among groups** of means. If we are interested only in testing the differences between all pairs of means, the Scheffe method is not the most sensitive procedure (Tukey is better). To make comparisons among groups of means, we will define a contrast as in Topic 4:

$$Q = \sum c_i \bar{Y}_i \text{ with } \sum c_i = 0 \text{ (or } \sum r_i c_i = 0 \text{ for unequal replication)}$$

We will reject the hypothesis ( $H_0$ ) that the contrast  $Q=0$  if the absolute value of  $Q$  is larger than a critical value  $F_S$ . This is the general form for Scheffe's F test:

$$\text{Critical value } F_S = \sqrt{df_{TR} F_{\alpha} (df_{TR}, df_{MSE})} \sqrt{MSE \sum \frac{c_i^2}{r_i}}$$

Note that previous pairwise comparisons are for the particular contrast 1 vs. -1. If we want to compare the control vs. the average of the three acids in Table 4.1, the contrast coefficients +3 -1 -1 -1 are multiplied by the means of the respective treatments.

$$Q = 4.190 * 3 + 3.868 * (-1) + 3.728 * (-1) + 3.640 * (-1) = 1.334$$

$$\text{Critical value } F_{S, 0.05, (3, 16)} = \sqrt{3 * 3.24} \sqrt{0.0086(3^2 + (-1)^2 + (-1)^2 + (-1)^2)/5} = 0.4479$$

The resulting  $|Q| > F_S$ , therefore we reject  $H_0$ . The average of the control (4.190-mg) is significantly different from the average of the three treatments with acid (3.745-mg). Remember that in these contrasts we are using **means** no **totals**.

### 5. 3. 2. Multiple-stage tests

The methods discussed so far provide simultaneous confidence intervals. By giving up the ability for simultaneous estimation with a single value, it is possible to obtain simultaneous tests with greater power using multiple-stage tests (MSTs). MSTs come in both step-up and step-down varieties but only step-down methods, which have been more widely used, are available in SAS. The best known MSTs are the Duncan and Student-Newman-Keuls (SNK) methods. Both use the studentized range statistic and, hence, are called **multiple range** tests. With means arrayed from the lowest to the highest, a multiple-range test gives significant ranges that become larger as the pairwise means to be compared are further apart in the array. Multiple range tests should only be used with balanced designs since they are inefficient with unbalanced ones.

The idea of step-down MSTs these tests is this: The more means (i.e. treatments) are compared, the smaller the probability that they are all the same. Therefore the maximum and minimum means are first compared pairwise using a significance level  $\gamma_t$ . If this  $H_0$  is accepted then the procedure stops, otherwise each subset of  $t-1$  means is tested against a significance  $\gamma_{t-1}$ . In other words, at the next stage, mean 1 is compared with mean  $t-1$ , and mean 2 with mean  $t$ . This process is repeated with smaller and smaller subsets down to 2. The EERC in a multiple range test is at most  $\gamma_t$ , and the CER is at most  $\gamma_2$ .

#### 5. 3. 2. 1. Duncan's multiple range tests

The test is identical to LSD for adjacent means in an array but requires progressively larger values for significance between means as they are more widely separated in the array. However for groups of two means uses the same value as LSD.

Duncan's test uses  $\gamma_n = 1 - (1 - \alpha)^{n-1}$ ,  $n = 2, \dots, t$ . For example, if  $\alpha = 0.05$  and  $t = 5$  then  $\gamma_5 = 0.18$ ,  $\gamma_4 = 0.14$ ,  $\gamma_3 = 0.098$ , and  $\gamma_2 = 0.05$ .

It controls the CER at the  $\alpha$  level but it has a high type I error rate (MEER). Its operating characteristics appear similar to those of Fisher's unprotected LSD at level  $\alpha$ . Since the last test is easier to compute, easier to explain, and applicable to unequal sample sizes, Duncan's method is not recommended by SAS. The higher power of Duncan's method compared to Tukey is due to its higher Type 1 error rate (Einot and Gabriel 1975). Duncan's test used to be the most popular method but many journals no longer accept it.

The procedure is to compute a set of critical values by using ST&D Table A7:

For Table 4.1 data

P	2	3	4
$q_{0.05(p, 16)}$	3.0	3.15	3.23
$R_p$	0.124	0.131	0.134

$$R_p = q_{\alpha, (p, \text{MSE df})} \sqrt{\frac{\text{MSE}}{r}}$$

### 5. 3. 2. 2. The Student-Newman-Keuls (SNK) test

Student-Newman-Keuls test uses  $\gamma_n = \alpha$ . This test is more conservative than Duncan's in that the type I error rate is smaller. The difference with Duncan is that SNK compares first the maximum and minimum means, and if that range is not significant, no further testing is done, and the set of means is declared homogeneous. Duncan continues the comparison using fixed levels for sets including different number of means. Because  $\alpha$  is generally lower, the power of SNK is generally lower than that of Duncan's test. It is often accepted by journals that do not accept Duncan's test.

If the range between the maximum and the minimum means is significant, SNK continues with the next comparison (between the minimum and the one before the maximum). If no difference is detected the test stops, if not it continues with the next mean..

The SNK test controls the EERC at the  $\alpha$  level but has poor behavior in terms of the EERP and MEER (Einot and Gabriel 1975). Consider ten population means that occur in five pairs such that means within pairs are equal, but there are large differences between pairs. All subset homogeneity hypotheses for three or more means are rejected. The SNK method then comes down to five independent tests, one for each pair, each at the  $\alpha$  level. The probability of at least one false rejection is :  $1 - (1 - 0.05)^5 = 0.23$ . As the number of means increases, the MEER approaches 1. Therefore, the SNK method is not recommended by SAS.

The procedure is to compute a set of critical values by using S&T Table A8. First compare the maximum and minimum means. If the range is not significant

$$W_p = q_{\alpha, (p, \text{MSE df})} \sqrt{\frac{MSE}{r}} \quad p = t, t-1, \dots, 2$$

For unequal r use the same correction as in Tukey (5. 3. 1. 3.). For Table 4.1 data:

p	2	3	4	
$q_{0.05 (p, 16)}$	3.0	3.65	4.05	Note that for $p = t$ $W_p =$ Tukey $w$
$W_p$	0.124	0.151	0.168	and for $p = 2$ $W_p =$ LSD

Table 5.9

Treatment	Mean	$W_p$	
Control	4.19		a
HCl	3.87		b
Propionic	3.73		c
Butyric	3.64		c

### 5. 3. 2. 3. The REGWQ method

A variety of MSTs that control MEER have been proposed, but these methods are not as well known as those of Duncan and SNK. An approach developed by Ryan, Einot and Gabriel, and Welsh (REGW) sets:

$$\gamma_p = 1 - (1 - \alpha)^{p/t} \quad \text{for } p < t-1 \quad \text{and} \quad \gamma_p = \alpha \quad \text{for } p \geq t-1.$$

The REGWQ method does the comparisons using a range test. This method appears to be among the most powerful step-down multiple range tests and is recommended by SAS for equal replication.

Assuming the sample means have been arranged in descending order from  $\bar{Y}_1$  through  $\bar{Y}_k$ , the homogeneity of means  $\bar{Y}_i, \dots, \bar{Y}_j$ , with  $i < j$ , is rejected by REGWQ if:

$$\bar{Y}_i - \bar{Y}_j \geq q(\gamma_p; p, \text{df}_{\text{MSE}}) \sqrt{\frac{MSE}{r}} \quad (\text{Use Table A.8 ST\&D})$$

For Table 5.1 data:

p	2	3	4
$\gamma_p$	0.025	0.05	0.05
$q \gamma_p (p, 16)$	3.49	3.65	4.05
Critical value	0.145	0.151	0.168

For  $p = t$  and  $p = t-1$  the critical value is as in SNK, but is larger for  $p < t-1$ . Note that the difference between HCl and propionic is significant with SNK but no significant with REGWQ ( $3.87 - 3.73 < 0.145$ ).

Table 5.10

Treatment	Mean	$F_s$
Control	4.19	a
HCl	3.87	b
Propionic	3.73	b c
Butyric	3.64	c

#### 5. 4. Conclusions and recommendations

There are at least twenty other parametric procedures available for multiple comparisons in addition to many non-parametric 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 an experiment-wise error rate (such as protected LSD and Scheffe's test). 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. The SAS manual makes the following additional recommendations: for controlling the MEER use the REGWQ method in a balanced design and Tukey method for unbalanced designs, which also gives confidence intervals.

One point to note is that unbalanced designs can give strange results. In the example of ST&D p 200, 4 treatments, A, B, C, and D have responses in the order  $A > B > C > D$ . A and D each have 2 replications while B and C each have 11. No significant difference was found between the extremes A and D but was detected between B and C.