Topic 5. Mean separation: Multiple comparisons [ST&D 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, we fail to reject H_0 and there is nothing more to do, except possibly redo the experiment, taking measures to make it more sensitive. If H_0 *is* rejected, then we conclude that at least one mean is significantly different from at least one other mean. The overall ANOVA gives no indication of which means are significantly different. If there are only two treatments, there is no problem; but if there are more than two treatments, the problem remains of needing to determine which means are significantly different. This is the process of mean separation.

Mean separation takes two general forms:

1. Planned, single degree of freedom F tests (orthogonal contrasts, last topic)
2. Multiple comparison tests that are suggested by the data (multiple comparison tests, Topic 5) itself (this topic).

Of these two methods, orthogonal F tests are preferred because they are more powerful than multiple comparison tests (i.e. they are more sensitive to differences than are multiple comparison tests). As you saw in the last topic, however, contrasts are not always appropriate because they must satisfy a number of strict constraints:

1. Contrasts are planned comparisons, so the researcher must have *a priori* knowledge about which comparisons are most interesting. This prior knowledge, in fact, determines the treatment structure of the experiment.

2. The set of contrasts must be orthogonal.

3. The researcher is limited to making, at most, (t – 1) comparisons.

Very often, however, there is no such prior knowledge. The treatment levels do not fall into meaningful groups, and the researcher is left with no choice but to carry out a sequence of multiple, unconstrained comparisons for the purpose of ranking and discriminating means. The different methods of multiple comparisons allow the researcher to do just that. There are many such methods, the details of which form the bulk of this topic, but generally speaking each involves 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

Selection of the most appropriate multiple comparison test is heavily influenced by the ***error rate***. Recall that a Type I error, occurs when one incorrectly rejects a true H_0. The **Type I error rate** is the fraction of times a Type I error is made. In a single comparison (imagine a simple *t* test) this is the value α. When comparing three or more treatment means, however, there are at least two different rates 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. In each experiment there is a total of 10 possible pairwise comparisons that can be made:

$$\text{Total possible pairwise comparisons (p)} = \frac{t(t-1)}{2}$$

For t = 5, p = (1/2)*(5*4) = 10
i.e. T_1 vs. T_2, T_3,T_4,T_5; T_2 vs. T_3,T_4,T_5; T_3 vs. T_4,T_5; T_4 vs. T_5

With 100 such experiments, therefore, there are a total of 1,000 possible pairwise comparisons. Suppose that there are no true differences among the treatments (i.e. H_0 is true) and that in each of the 100 experiments, one Type I error is made. Then the CER over all experiments is:

CER= (100 mistakes) / (1000 comparisons) = 0.1 or 10%

The EER is

EER= (100 experiments with mistakes) / (100 experiments) = 1 or 100%.

The EER is the probability of making at least one Type I error in the experiment. As the number of means (and therefore the number of possible comparisons) increases, the chance of making at least one Type I error approaches 1. To preserve a low experiment-wise error rate, then, the comparison-wise error rate must be held extremely low. Conversely, to maintain a reasonable comparison-wise error rate, the experiment-wise error rate will inflate.

The relative importance of controlling these two Type I error rates depends on the objectives of the study, and different multiple comparison procedures have been developed based on different philosophies of controlling these two kinds of error. In situations where incorrectly rejecting one comparison may jeopardize the entire experiment or where the consequence of incorrectly rejecting one comparison is as serious as incorrectly rejecting a number of comparisons, the control of experiment-wise error rate is more important. On the other hand, when one erroneous conclusion will not affect other inferences in an experiment, the comparison-wise error rate is more pertinent.

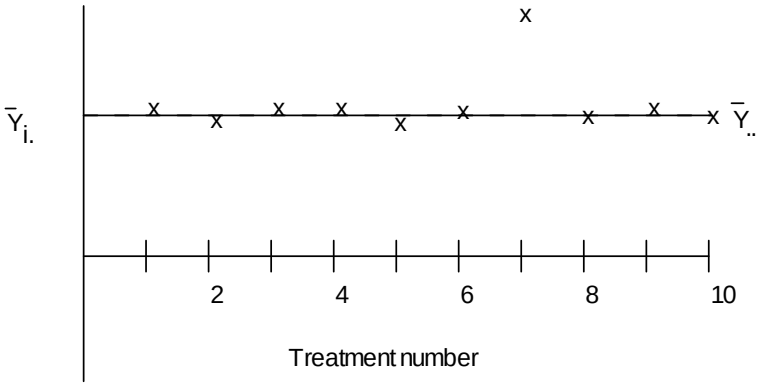
The experiment-wise error rate is always larger than the comparison-wise error rate. It is difficult to compute the exact experiment-wise error rate because, for a given data set,

Type I errors are not independent. But it is possible to compute an upper bound for the EER by assuming that the probability of a Type I error for any single comparison is α and is independent of all other comparisons. In that case:

Upper bound EER = $1 - (1 - \alpha)^p$ where $p = \frac{t(t-1)}{2}$, as before.

So for 10 treatments and $\alpha = 0.05$, the upper bound of the EER is 0.9 (EER = $1 - (1 - 0.05)^{45} = 0.90$ or 90%).

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. Such a situation is indicated graphically below:



A simple ANOVA will probably reject H_0, so the experimenter will want to determine which specific means are different. Even though one mean is truly different, there is still a chance of making a Type I error in each pairwise comparison among the 9 similar treatments. An upper bound on this probability is computed by setting t = 9 in the above formula, giving a result of 0.84. That is, the experimenter will incorrectly conclude that two truly similar effects are actually different **84% of the time**. This is called the experiment-wise error rate under a partial null hypothesis, the partial null hypothesis in this case being that the subset of nine treatment means are all equal to one another.

So we can distinguish between the EER under the complete null hypothesis, in which all treatment 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 the following four categories:

- ***CER*** = comparison-wise error rate
- ***EERC*** = experiment-wise error rate under a complete null hypothesis (standard EER)
- ***EERP*** = experiment-wise error rate under a partial null hypothesis.
- ***MEER*** = maximum experiment-wise error rate under any complete or partial null hypothesis.

5. 3. Multiple comparisons tests

Statistical methods for making two or more inferences while controlling the Type I error rates (CER, EERC, EERP, MEER) are called ***simultaneous inference methods***. The material in this sectionis based primarily on ST&D chapter 8 and on the SAS/STAT manual (GLM Procedure). The basic techniques of multiple comparisons fall into two groups:

1. Fixed-range tests: Those which provide confidence intervals and tests of hypotheses.
2. Multiple-range tests: Those which provide only tests of hypotheses.

To illustrate the various procedures, we will use the data from two different experiments given in Table 4-1 (previous class, equal replication) and 5-1 (below, unequal replication). The ANOVAs 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 a single range for making all possible pairwise comparisons in experiments with equal replications across treatment groups (i.e. in balanced designs). Many fixed-range procedures are available, and considerable controversy exists as to which procedure is most appropriate. We will present four commonly used procedures, moving from the less conservative to the more conservative: LSD, Dunnett, Tukey, and Scheffe. Other pairwise tests are discussed in the SAS manual.

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

One of the oldest, simplest, and most widely misused multiple pairwise comparison tests is the least significant difference (LSD) test. The LSD is based on the t-test (ST&D 101); in fact, it is simply a sequence of many t-tests. Recall the formula for the t statistic:

$$t = \frac{\bar{Y}_{(r)} - \mu}{s_{\bar{Y}(r)}} \quad \text{where} \quad s_{\bar{Y}(r)} = \frac{s}{\sqrt{r}}$$

This t statistic is distributed according to a t distribution with (r – 1) degrees of freedom. The LSD test declares the difference between means $\overline{Y}_i$ and $\overline{Y}_j$ of treatments i and j to be significant when:

$$|\overline{Y}_i - \overline{Y}_j| > \text{LSD, where}$$

$$LSD = t_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\left(\frac{1}{r_1}+\frac{1}{r_2}\right)}$$ for unequal r (SAS calls this a repeated t test)

$$LSD = t_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\frac{2}{r}}$$ for equal r (SAS calls this an 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. As an example, here are the calculations for Table 4.1. Note that the significance level selected for pairwise comparisons does not have to conform to the significance level of the overall F test. To compare procedures across the examples to come, we will use a common $\alpha = 0.05$.

From Table 4-1, MSE = 0.0086 with 16 df.

$$LSD = t_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\frac{2}{r}} = 2.120\sqrt{0.0086\frac{2}{5}} = 0.1243$$

So, if the absolute difference between any two treatment means is more than 0.1243, the treatments are said to be significantly different at the 5% confidence level. As the number of treatments increases, it becomes more and more difficult, just from a logistical point of view, to identify those pairs of treatments that are significantly different. A systematic procedure for comparison and ranking begins by arranging the means in descending or ascending order as shown below:

Control	4.19
HCl	3.87
Propionic	3.73
Butyric	3.64

Once the means are so arranged, compare the largest with the smallest mean. If these two means are significantly different, compare the next largest mean with the smallest. Repeat this process until a non-significant difference is found. Label these two and any means in between with a common lower case letter by each mean. Repeat the process with the next smallest mean, etc. Ultimately, you will arrive at a mean separation table like the one shown below:

Table 5.5

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

Pairs of treatments that are not significantly different from one another share the same letter. For the above example, we draw the following conclusions at the 5% confidence level:

- All acids reduced shoot growth.
- The reduction was more severe with butyric and propionic acid than with HC1.
- We do not have evidence to conclude that propionic acid is different in its effect than butyric acid.

When all the treatments are equally replicated, note that only one LSD value is required to test all six possible pairwise comparisons between treatment means. This is not true in cases of unequal replication, where different LSD values must be calculated for each comparison involving different numbers of replications.

For the second data set (Table 5.1.), we find the 5% LSD for comparing the control with Feed B to be:

$$LSD = t_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\left(\frac{1}{r_1}+\frac{1}{r_2}\right)} = 2.074\sqrt{0.00224\left(\frac{1}{6}+\frac{1}{5}\right)} = 0.0594$$

The other required LSD's are:

A vs. Control = 0.0531
A vs. B = 0.0560
A vs. C = 0.0509
B vs. C= 0.0575
C vs. Control = 0.0546

Using these values, we can construct a mean separation table:

Treatment	**Mean**	**LSD**
Feed B	1.45	a
Feed A	1.36	b
Feed C	1.33	b
Control	1.20	c

Thus, at the 5% level, we conclude all feeds cause significantly greater weight gain than the control. Feed B causes the highest weight gain; Feeds A and C are equally effective.

One advantage of the LSD procedure is its ease of application. Additionally, it is easily used to construct confidence intervals for mean differences. The 1- α confidence limits of the quantity (μ_A - μ_B) are given by=

$$(1 - \alpha)\ \text{CI for } (\mu_A - \mu_B) = (\bar{Y}_A - \bar{Y}_B) \pm \text{LSD}$$

Because fewer comparisons are involved, 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. The test is primarily intended for use when there is no predetermined structure to the treatments. If a large number of means are to be compared and the ones compared are selected *after* the ANOVA and the comparisons target those means with most different values, the actual error rate will be much higher than predicted.

The LSD test is the only test for which the comparison-wise error rate equals α. This is often regarded as too liberal (i.e. too ready to reject H_0). It has been suggested that the EEER can be maintained at α by performing the overall ANOVA test at the α level and making further comparisons *if and 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. In those cases, a preliminary F test controls only the EERC, not the EERP.

5. 3. 1. 2. Dunnett's Method

In certain experiments, one may desire only to compare a control with each of the other treatments, such as comparing a standard variety or chemical with several new ones. Dunnett's method performs such an analysis while holding the maximum experimentwise error rate under any complete or partial null hypothesis (MEER) to a level not exceeding the stated α.

In this method, a t* value is calculated for each comparison. This tabular t* value for determining statistical significance, however, is not the Student's t but a special t* given in Appendix Tables A-9a and A-9b (ST&D p 624-625). Let $\bar{Y}_0$ represent the control mean with r_0 replications, then:

$$DLSD = t^*_{\frac{\alpha}{2}, df_{MSE}} \sqrt{MSE\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}$$ for unequal r ($r_0 \neq r_i$)

and

$$DLSD = t^*_{\frac{\alpha}{2}, df_{MSE}} \sqrt{MSE\frac{2}{r}}$$ for equal r ($r_0 = r_i$)

From the seed treatment experiment in Table 4-1, MSE = 0.0086 with 16 df and the number of comparisons (p)= 3.

By Table A-19b, $t^*_{\frac{\alpha}{2},16} = 2.59$.

$$DLSD = t^*_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\frac{2}{r}} = 2.59\sqrt{0.0086\frac{2}{5}} = 0.1519$$

(**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:

$(1 - \alpha)$ CI for $(\mu_0 - \mu_i) = (\bar{Y}_0 - \bar{Y}_i) \pm$ DLSD

The limits of these differences are,

Control	-	Butyric	=	0.32 ± 0.15
Control	-	HC1	=	0.46 ± 0.15
Control	-	Propionic	=	0.55 ± 0.15

We have 95% confidence that the 3 ranges will include **simultaneously** the true differences.

When treatments are not equally replicated, as in the feed ration experiment, there are different DLSD values for each of the comparisons. To compare the control with Feed-C, first note that $t^*_{0.025,22} = 2.517$ (from SAS; by Table A-9b, t* is 2.54 or 2.51 for 20 and 24 df, respectively):

$$DLSD = t^*_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\left(\frac{1}{r_0}+\frac{1}{r_1}\right)} = 2.517\sqrt{0.00224\left(\frac{1}{6}+\frac{1}{7}\right)} = 0.0663$$

Since $|\bar{Y}_0 - \bar{Y}_C| = 0.125$ is larger than 0.06627, the difference is significant. All other differences with the control, being larger than this, are also significant.

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

Tukey's test was designed specifically for **pairwise comparisons**. This test, sometimes called the "honestly significant difference test" (HSD), controls the MEER when the sample sizes are equal. Instead of ***t*** or ***t****, it uses the statistic $q_{\alpha,p,df_{MSE}}$ that is obtained from Table A-8. The Tukey critical values are larger than those of Dunnett because the

Tukey family of contrasts is larger (all possible pairs of means instead of just comparisons to a control). The critical difference in this method is labeled *w*:

$$w = q_{\alpha,p,df_{MSE}}\sqrt{\frac{MSE}{2}\left(\frac{1}{r_1}+\frac{1}{r_2}\right)}$$ for unequal r

$$w = q_{\alpha,p,df_{MSE}}\sqrt{\frac{MSE}{r}}$$ for equal r

Aside from the new critical value, things looks basically the same as before, except notice that here we do not multiply MSE by a factor of 2 *because Table A-8 already includes the factor √2 in its values*. For example, for p = 2, df = ∞ (equivalent to the standard normal distribution Z), and α= 5%, the critical value is 2.77, which is equal to 1.96 * √2.

Considering the seed treatment data (Table 4.1): $q_{0.05,(4,16)} = 4.05$; and:

$$w = q_{\alpha,(p,df_{MSE})}\sqrt{\frac{MSE}{r}} = 4.05\sqrt{\frac{0.0086}{5}} = 0.1680$$

(**Note that w = 0.1680 > DLSD = 0.1519 > LSD = 0.1243**)

By this method, the means separation table looks like:

Table 4.1

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

Like the LSD and Dunnett's methods, this test detects significant differences between the control and all other treatments. But unlike with the LSD method, it detects no significant differences between the HCl and Propionic treatments (compare with Table 5.5). This reflects the lower power of this test.

For unequal r, as in the feeding experiment in Table 5.3, the contrast between the Control with Feed-C would be tested using:

$q_{0.05,(4,22)} = 3.93$ $w = 3.93\sqrt{0.00224(\frac{1}{6}+\frac{1}{7})/2} = 0.0732$

Since $|\overline{Y}_{Cont} - \overline{Y}_C| = 0.125$ is larger than 0.0731, it is significant. As in the LSD, the only pairwise comparison that is not significant is that between Feed C ($Y_C = 1.330$) and Feed A ($Y_A = 1.361$).

5. 3. 1. 4. Scheffe's F test for pairwise comparisons

Scheffe's test is compatible with the overall ANOVA *F* test in the sense 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. This includes ***all possible pairwise and group comparisons***. Since this procedure controls MEER while allowing for a larger number of comparisons, it is less sensitive (i.e. more conservative) than other multiple comparison procedures.

The Scheffe critical difference (SCD) has a similar structure as that described for previous tests, scaling the critical F value for its statistic:

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

$$SCD = \sqrt{df_{Trt} F_{\alpha, df_{Trt}, df_{MSE}}} \sqrt{MSE \frac{2}{r}} \quad \text{for equal r}$$

For the seed treatment data (Table 4-1), MSE = 0.0086 with df_{TR}= 3, df_{MSE}= 16, and r= 5

$$SCD_{0.05} = \sqrt{3 * 3.24} \sqrt{0.0086 \frac{2}{5}} = 0.1829$$

(**Note that SCD = 0.1829 > w = 0.1680 > DLSD = 0.1519 > LSD = 0.1243**)

Again, if the difference between a pair of means is greater than SCD, that difference will be declared significant at the given α level, *while holding MEER below α.* The table of means separations:

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 based on unequal replications, a different SCD is required for each comparison. For the animal feed experiment, critical difference for the contrast between the Control and Feed-C is:

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

Since $|\overline{Y}_{Cont} - \overline{Y}_C| = 0.125$ is larger than 0.0796, it is significant. Scheffe's procedure is also readily used for interval estimation:

(1 – α) CI for (μ_0 - μ_i) = ($\overline{Y}_0 - \overline{Y}_i$) ± SCD

The resulting intervals are **simultaneous** in that the probability is at least (1 – α) that all of them are true simultaneously.

5.3.1.5 Scheffe's F test for group comparisons

The most important use of Scheffe's test is for arbitrary **comparisons among groups of means**. We use the word "arbitrary" here because, unlike the group comparisons using contrasts, group comparisons using Scheffe's test do not have to be orthogonal, nor are they limited to (t – 1) questions. If you are interested only in testing the differences between all pairs of means, the Scheffe method is not the best choice; Tukey's is better because it is more sensitive while controlling MEER. But if you want to "mine" your data by making all possible comparisons (pairwise and group comparisons) while still controlling MEER, Scheffe's is the way to go.

To make comparisons among groups of means, you first define a contrast, as in Topic 4:

Q= $\sum c_i \ \bar{Y}_{i.}$ with the constraint that $\sum c_i = 0$ (or $\sum r_i \ c_i = 0$ for unequal r)

We will reject the null 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 test:

Critical value $F_S = \sqrt{df_{Trt} F_{\alpha, df_{Trt}, df_{MSE}}} \sqrt{MSE \sum_{i=1}^{t} \frac{c_i^2}{r_i}}$

Note that the previous expressions for Scheffe pairwise comparisons (5.3.1.4.) are for the particular contrast 1 vs. -1. If we want to compare the control to the average of three acid treatments in Table 4.1, the contrast coefficients are +3 -1 -1 –1.

In this case Q is calculated by multiplying the coefficients for the means of the respective treatments.

$$Q = \sum_{i=1}^{t} c_i \bar{Y}_i = 4.190(3) + 3.868(-1) + 3.728(-1) + 3.640(-1) = 1.334$$

The critical value $F_{S\ 0.05,\ (3,\ 16)}$ value for this contrast is:

$$F_S = \sqrt{df_{Trt} F_{\alpha, df_{Trt}, df_{MSE}}} \sqrt{MSE \sum_{i=1}^{t} \frac{c_i^2}{r_i}} = \sqrt{3(3.24)0.0086 \frac{3^2 + (-1)^2 + (-1)^2 + (-1)^2}{5}} = 0.4479$$

Since $|Q| = 1.334 > 0.4479 = F_S$, we reject H_0. The average of the control (4.190 mg) is significantly different from the average of the three acid treatments (3.745 mg).

Again, with Scheffe's method, you can test any conceivable set of contrasts, even if they number more than (t – 1) questions and are not orthogonal. The price you pay for this freedom, however, is very low sensitivity. Scheffe's is the most conservative method of comparing means; so if Scheffe's declares a difference to be real, you can believe it.

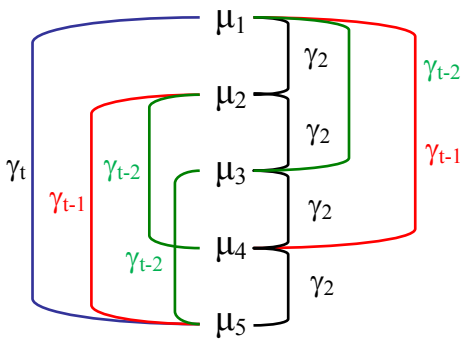
Remember that in these contrasts we are using **means** no **totals**.

5. 3. 2. Multiple-stage tests

Before we start: Multiple range tests should only be used with balanced designs since they are inefficient with unbalanced ones.

The methods discussed so far are all "fixed-range" tests, so called because they use a single, fixed value to test hypotheses and build simultaneous confidence intervals. If one forfeits the ability to build simultaneous confidence intervals with a single value, it is possible to obtain simultaneous hypothesis tests of greater power using multiple-stage tests (MSTs). MSTs come in both step-up (first comparing closest means, then more distant means) and step-down varieties (the reverse); but only step-down methods, which are more widely used, are available in SAS. The best known MSTs are the Duncan and the Student-Newman-Keuls (SNK) methods. Both use the *studentized range statistic (q)* and, hence, also go by the name **multiple range** tests. With means arranged from the lowest to the highest, a multiple-range test provides critical distances or ranges that become smaller as the pairwise means to be compared become closer together in the array. Such a strategy allows the researcher to allocate test sensitivity where it is most needed, in discriminating neighboring means.

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. The general strategy is: First, the maximum and minimum means are compared pairwise using the largest critical value since the comparison involves all the means. If this H_0 is accepted, the procedure stops. Otherwise, the analysis continues by comparing pairwise the two sets of next-most-extreme means (i.e. μ_1 vs. μ_{t-1}, and μ_2 vs. μ_t) using a smaller critical value, because groups are now smaller. This process is repeated with closer and closer pairs of means until one reaches the set of (t – 1) pairs of adjacent means, compared pairwise using the smallest critical value. The larger the range of the ranks, the larger the tabled critical point.



Graphical depiction of the general strategy of step-down MSTs. In this figure, the means are arranged highest (μ_1) to lowest (μ_5). The significance levels of each of the 10 possible pairwise comparisons are indicated by the γ.

5. 3. 2. 1. Duncan's multiple range tests (Table A-7)

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.

It controls the CER at the α 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 α. 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, in fact, 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.

To compute Duncan critical ranges (R_p), use the following expression, plugging in the appropriate values of the Studentized range statistic (q_α):

$$R_p = q_{\alpha_{p-1},p,df_{MSE}}\sqrt{\frac{MSE}{r}}$$

The procedure is to compute a set of critical values by using ST&D Table A-7.:

For the seed treatment data in Table 4.1:

P	2	3	4
$q_{0.05\,(p,\,16)}$	3.00	3.15	3.23
R_p	0.124	0.131	0.134

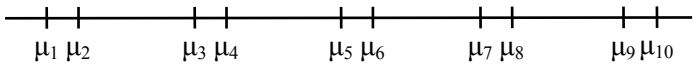
Note that the critical difference for p=2 is the same as the LSD test!

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

Student-Newman-Keuls test (SNK) is more conservative than Duncan's in that the Ttype I error rate is smaller. This is because SNK simply uses α as the significance level at all stages of testing, again stopping the analysis at the highest level of non-significance. Because α is lower than Duncan's variable significance values, the power of SNK is generally lower than that of Duncan's test.

SNK is often accepted by journals that do not accept Duncan's test.

The SNK test controls the EERC at the α level but it behaves poorly in terms of the EERP and MEER (Einot and Gabriel 1975). To see this consider ten population means that cluster in five pairs such that means within pairs are equal but there are large differences between pairs:



In such case, 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 conducted at the α 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 since it does not control well the maximum experiment wise error rate under any partial null hypothesis (e.g. the one in the figure above).

The procedure is to compute a set of critical values by using ST&D Table A-8. First compare the maximum and minimum means. If the range is not significant

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

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

For Table 4.1 data:

p	2	3	4
$q_{0.05\ (p,\ 16)}$	3.00	3.65	4.05
W_p	0.124	0.151	0.168

Note that for p= t W_p= Tukey *w* 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:

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

The REGWQ method performs 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 (i.e. balanced design).

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

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

For Table 5.1 data:

p	2	3	4
γ_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, not to mention the many non-parametric and multivariate methods. There is no consensus as to which is the most appropriate procedure to recommend to all users. One main difficulty in comparing the procedures is the different kinds of Type I error rates used, namely, experiment-wise versus comparison-wise. All this is to say that the difference in performance of any two procedures is likely due to the different underlying philosophies of Type I error control than to the specific techniques used.

To a large extent, the choice of a procedure is subjective and hinges on a choice between a comparison-wise error rate (such as LSD) and an experiment-wise error rate (such as Tukey and Scheffe's test).

Some suggested rules of thumb:

1. When in doubt, use Tukey. Tukey's method is a good general technique for carrying out all pairwise comparisons, enabling you to rank means and put them into significance groups, while controlling MEER.

2. Use Dunnett's (more powerful than Tukey's) if you only wish to compare each treatment level to a control.

3. Use Scheffe's if you wish to test a set of non-orthogonal group comparisons *OR* if you wish to carry out group comparisons *in addition to* all possible pairwise comparisons. MEER will be controlled in both cases.

The SAS manual makes the following additional recommendation: For controlling MEER for all pairwise comparisons, use REGWQ for balanced designs and Tukey for unbalanced designs.

One final point to note is that severely unbalanced designs can yield very strange results, regardless of means separation method. To illustrate this, consider the example on page 200 of ST&D. In this example, an experiment with four 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. The strange result: The extreme means (A and D) are found to be not significantly different, but the intermediate means (B and C) are.