

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

5.1. Basic concepts

If there are more than 2 treatments the problem is to determine which means are significantly different. This process can take 2 forms.

- * **Planned F test** (Orthogonal F test, Topic 4).
 - * Planned
 - * Independent
 - * More precise and powerful
- * **Effects suggested by the data** (multiple comparison tests, Topic 5)
 - No preliminary grouping is known
 - Useful in experiments where no particular relationships among 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 α .

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

Example

Suppose the experimenter conducts **100 experiments** with **5 treatments** each.

In each experiment there are 10 possible pairwise comparisons: $(t(t-1)/2)$

There are a total of **1000 possible comparisons**.

Suppose that there are **no true differences among the treatments**.

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

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

Relationship between CER and EER

- **EER** is the probability of a Type I error **for the complete experiment**. As the number of means increases, the chance of making at least one Type I error approaches 1.
- To preserve a low **EER**, the **CER** has to be kept very low. Conversely, to maintain a reasonable **CER**, the **EER** must be very large.
- The relative importance of controlling these two Type I error ratios depends on the objectives of the study and the number of treatments.
- Different multiple comparison procedures have been developed based on different philosophies of controlling these two kinds of errors.
- There is no universal criterion that enables us to decide whether a **CER** or an **EER** is more appropriate to be controlled.
- When 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 **EER** is most important.
- On the other hand, when one erroneous conclusion does not affect the remaining inferences in an experiment, the **CER** is pertinent.

COMPUTATION

It is difficult to compute the **EER** 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 α and is independent of the other comparison probabilities. In this case:

$$\text{Upper bound EER} = 1 - (1 - \alpha)^p$$

$p = N^{\circ}$ of pairs to compare = $t(t-1)/2$.

$(1 - \alpha)$ = Pb of not making an error in one comparison

$(1 - \alpha)^p$ = Pb of not making an error in p comparisons

Example

If 10 treatments $\Rightarrow p = 10 * 9 / 2 = 45$ pairwise comparisons

If $\alpha = 0.05 \Rightarrow$ pb. of not making a Type 1 error in one comparison = 0.95

Pb. of not making a Type 1 error in p comparisons: $(1 - \alpha)^p = 0.95^{45} = 0.1$

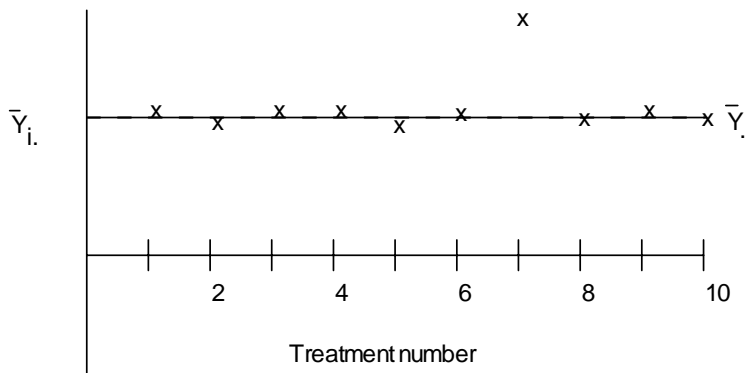
Probability of making 1 or more Type 1 error in p comparisons: $1 - 0.1 = 0.9$

Upper bound **EER** = 0.9.

This formula can be used to fix **EER** and then calculate the required α

$$\text{EER} = 0.1 \Rightarrow (1 - \alpha)^{45} = 1 - 0.1 \Rightarrow (1 - \alpha) = 0.9^{1/45} = 0.997 \Rightarrow \alpha = 0.002$$

Partial null hypothesis: Suppose there are 10 treatments and one shows a significant effect while the other 9 are approximately equal. ANOVA will reject H_0 .



There is a probability of making a Type I error between each of the 9 similar treatments

The upper bound **EER** is computed by setting $t = 9$ in the above formula.

It is 0.84, which is therefore the **EER under a partial null hypothesis**.

That is, the experimenter will incorrectly conclude that some pair of similar effects are different 84% of the time.

Examples

Table 4.1. Results (mg dry weight) of an experiment (CRD) to determine the effect of seed treatment by acids on the early growth of rice seedlings.

Treatment	Replications					Total	Mean
						$Y_{i.}$	$\bar{Y}_{i.}$
Control	4.23	4.38	4.1	3.99	4.25	20.95	4.19
HCl	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						$Y_{..} = 77.13$	$\bar{Y}_{..} = 3.86$

Table 4.2. ANOVA of data in Table 4.1.

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

Treat.								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	

Complete and partial null hypothesis in SAS

EER under the complete null hypothesis: all population means are equal

EER under a partial null hypothesis: some means are equal but some differ.

SAS subdivides the error rates into:

CER = comparison-wise error rate

EERC = experiment-wise error rate under complete null hypothesis (the 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 comparison tests ST&D Ch. 8 and SAS/STAT (GLM)

Simultaneous inference methods: Statistical methods for making two or more inferences while controlling the probability of making at least one Type I error.

Multiple comparison techniques divide themselves into two groups:

- **Fixed-range tests:** Those which can provide confidence intervals and tests of hypothesis. **One range** for testing all differences **in balanced** designs
- **Multiple-stage tests:** Those which are essentially only tests of hypothesis. **Variable ranges.**

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**
 - **Dunnnett**
 - **Tukey**
 - **Scheffe.**

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

LSD test is one of the simplest and one of the **most widely misused**.

LSD test declares the \neq between means \bar{Y}_i and \bar{Y}_j to be significant when:

$$|\bar{Y}_i - \bar{Y}_j| > t_{\alpha/2, \text{MSE df}} \sqrt{\text{MSE} \frac{2}{r}} \text{ for equal } r \text{ (SAS: LSD test)}$$

Where **MSE** = pooled s^2 was calculated by PROC ANOVA or PROC GLM.

2 is coming from the $X_1 - X_2 \sim t(\mu_1 - \mu_2, s_1^2 + s_2^2)$, if $s_1^2 = s_2^2 \Rightarrow s_1^2 + s_2^2 = 2 s_p^2$

From **Table 4.1**: $\alpha = 0.05$, $\text{MSE} = 0.0086$ with 16 df.

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

If $|\bar{Y}_i - \bar{Y}_j| > 0.1243$ or more, the treatments are said to be significantly different at the 5% level.

A systematic procedure of comparison is to arrange the means in descending or ascending order as shown below.

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

- 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.

When all the treatments are **equally replicated**, only **one LSD** value is required to test all possible comparisons. One advantage of the LSD procedure is its ease of application.

LSD is readily used to construct confidence intervals for mean differences.

The $1 - \alpha$ confidence limits are = $\bar{Y}_A - \bar{Y}_B \pm \text{LSD}$

LSD with different number of replications

Different LSD must be calculated for each comparison involving different numbers of replications.

$$|\bar{Y}_i - \bar{Y}_{i'}| > t_{\alpha/2, \text{MSE df}} \sqrt{MSE \left(\frac{1}{r_1} + \frac{1}{r_2} \right)} \text{ for unequal } r \text{ (SAS: repeated t test)}$$

For **Table 5.1** with unequal replications:

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 vs. Feed-B** (1.45-1.20=0.25) is,

$$\text{LSD}_{0.025} = 2.074 \sqrt{0.00224 \left(\frac{1}{6} + \frac{1}{5} \right)} = 0.0595, \text{ and } |\bar{Y}_i - \bar{Y}_{i'}| > 0.0595 \text{ Reject } H_0$$

Because of the **unequal replication**, the **LSD** values and the length of the **confidence intervals vary** among different pairs of mean comparisons

General considerations

- The LSD test is much safer when the means to be compared are selected *in advance* of the experiment.
- It is primarily intended for use when there is no predetermined structure to the treatments (e.g. in variety trials).
- 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 α level by performing the overall ANOVA test at the α 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

Compare a **control** with each of several other treatments

- Dunnett's test holds the maximum **EER** under any complete or partial null hypothesis (**MEER**) $< \alpha$.
- In this method a t^* value is calculated for each comparison.
- The tabular t^* value is given in **Table A-9 (ST&D p625)**.

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

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

From Table 4-1, MSE = 0.0086 with 16 df and $p=3$, $t^*_{\alpha/2}=2.59$ (Table A9)

$$DLSD_{0.025} = 2.59 \sqrt{0.0086 \frac{2}{5}} = 0.152 \quad \textbf{DLSD= 0.152 > LSD= 0.124!}$$

0.152= least significant difference between control and any other treatment.

Control - HC1 = 4.19 - 3.87 = 0.32. Since 0.32 > 0.152 (DLSD) \Rightarrow 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 ± 0.15
Control	-	HC1	=	0.46 ± 0.15
Control	-	propionic	=	0.55 ± 0.15

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

For **unequal replication**: **Table 5-1**

To compare **control** with feed-C, $t^*_{0.025, 22, p=3} = 2.517$ (from SAS)

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

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 for **all possible pairwise comparisons**.

The test is sometimes called "honestly significant difference test" HSDT

It 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** (distribution $(\bar{Y}_{MAX} - \bar{Y}_{MIN})/s_{\bar{Y}}$)

$$w = q_{\alpha,(p, \text{MSE df})} \sqrt{\frac{MSE}{r}} \quad \text{for equal } r \quad (\sqrt{2} \text{ is missing because is inside } q)$$

$$p=2 \quad \text{df}=\infty \quad \alpha=0.05 \quad q=2.77 = 1.96 * \sqrt{2}$$

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

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

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

Tukey critical value is larger than that of Dunnett because the Tukey family of contrasts is larger (all pairs of means).

Table 4.1

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

For **unequal r**, as in **Table 5.1**, 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.

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

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

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

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

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

Unequal replications: a different SCD is required for each comparison. The contrast **Control** vs. Feed-C (Table 5-1),

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

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**. The probability is at least $1 - \alpha$ that all of them are simultaneously true.

Scheffe's comparisons among groups

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, Tukey is more sensitive than Scheffe.

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:

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

In the previous pairwise comparisons the contrast is 1 vs. -1 $\Rightarrow \sum \frac{c_i^2}{r_i} = 2/r$.

Example

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**.

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

Critical value:

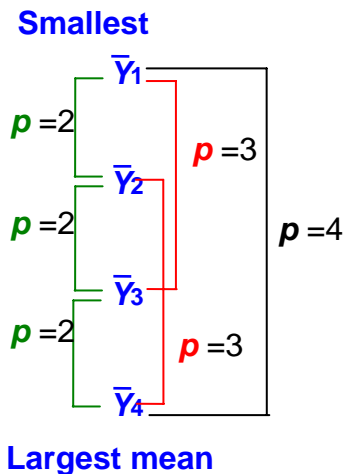
$$F_{S, 0.05, (3, 16)} = \sqrt{3 * 3.24} \sqrt{0.0086(3^2 + (-1)^2 + (-1)^2 + (-1)^2)/5} = 0.4479$$

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

5.3.2. Multiple-stage tests

- By giving up the facility for simultaneous estimation with one value, it is possible to obtain tests with greater power: **multiple-stage** tests (MSTs).
- The best known MSTs are: **Duncan** and **Student-Newman-Keuls (SNK)**
- They use **studentized range statistic**.

Multiple range tests should **only be used with balanced designs** since they are inefficient with unbalanced ones.



- With means arrayed from the lowest to the highest, a multiple-range test gives significant ranges that become smaller as the pairwise means to be compared are closer in the array
- Imagine a Tukey test where you 1st test the most distant means and then **ignore them** in the next level, resulting in a smaller **p** and smaller critical value in **Table 8**

5.3.2.1. Duncan's multiple range tests: no longer accepted

- 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.
- It controls the CER at the α level but it has **a high type I error rate (MEER)**.
- Duncan's test used to be the most popular method but many journals no longer accept it and is not recommended by SAS.

5.3.2.2. The Student-Newman-Keuls (SNK) test

- This test is more conservative than Duncan's in that the type I error rate is smaller.
- It is often accepted by journals that do not accept Duncan's test.
- The SNK test **controls** the **EERC** at the α level
- ...but has **poor** behavior in terms of the **EERP** and **MEER**. This method is not recommended by SAS.

The procedure is to compute a set of critical values (ST&D **Table A8**).

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

For Table 4.1

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

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

- Remember that HCl vs. Propionic was NS in Tukey's procedure.
- SNK is more sensitive than Tukey
- But...the cost is larger **EERP**!

Assume the following **partial null hypothesis** (means from smallest to largest)

$$\underline{\bar{Y}_1 = \bar{Y}_2} \neq \underline{\bar{Y}_3 = \bar{Y}_4} \neq \underline{\bar{Y}_5 = \bar{Y}_6} \neq \underline{\bar{Y}_7 = \bar{Y}_8} \neq \underline{\bar{Y}_9 = \bar{Y}_{10}}$$

There are 5 NS independent pairs compared by LSD \Rightarrow

$$\text{EERP} = 1 - (1 - 0.05)^5 = 0.23 \text{ Higher EERP than Tukey!}$$

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.

Ryan, Einot and Gabriel, and Welsh (**REGWQ**) method:

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

In Table 4.1 $p < 3$ $p \geq 3$

The REGWQ method does the comparisons using a range test.

This method appears to be among the most powerful multiple range tests and is recommended by SAS **for equal replication**.

Assuming the sample means have been arranged in order from \bar{Y}_1 through \bar{Y}_k , the homogeneity of means $\bar{Y}_i, \dots, \bar{Y}_j$, 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 4.1** data:

p	2	3	4
γ_p	0.0253	0.05	0.05
$q \gamma_p (p, 16)$	3.49	3.65	4.05
Critical value	0.145	0.151	0.168
	>SNK	=SNK	=SNK

$$\gamma_2 = 1 - (1 - \alpha)^{p/t} = 1 - 0.95^{2/4} = \mathbf{0.0253} \text{ (not in Table 8, use SAS)}$$

For $p < t-1$ the REGWQ critical value (0.145) is larger than SNK (0.124).

Note that the difference between HCl and propionic is significant with SNK but no significant with REGWQ ($3.87 - 3.73 < 0.145$).

Treatment	Mean	F_5
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. **The price of the better EERP is a lower sensitivity**

5. 4. Conclusions and recommendations

- There are **many procedures available for multiple comparisons**. At least 20 other parametric and 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 results from the **different kinds of Type I error rates used**, namely, experiment-wise versus comparison-wise.
- 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 less appropriate than **Tukey's** test, 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**.
- ST&D p 200: 4 treatments, A, B, C, and D with $A > B > C > D$. 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 (LSD).

Treatment	Data	Mean	
A	10, 17	13.5	* NS
B	10, 11, 12, 12, 13, 14, 15, 16, 16, 17, 18	14.0	
C	14, 15, 16, 16, 17, 18, 19, 20, 20, 21, 22	18.0	
D	16,21	18.5	