Pairwise multiple comparisons of treatment means in agronomic research[1]

Samuel G. Carmer and William M. Walker[2]

ABSTRACT

Pairwise multiple comparisons of treatment means are appropriate in the statistical analysis of some agronomic experiments. This paper includes a review and definitions of the types and frequency rates of statistical errors with regard to pairwise multiple comparisons. Of the 10 pairwise multiple comparisons procedures described herein, the least significant difference is the procedure of choice when the appropriate contrasts among treatments each involve only two of the treatment means. This choice is based on considerations of error rates, power, and correct decision rates as well as simplicity of computation.

Additional index words: **Duncan's multiple range test, Least significant difference, Statistical analysis, Waller-Duncan k-ratio t test.**

THE OBJECTIVE of a well-designed experiment is to answer questions of concern to the experimenter. The most appropriate and most informative method of statistical analysis of the data will be that procedure which provides the best answers to those questions. Most designed experiments include treatments selected for the purpose of answering specific questions. Frequently these specific questions are best answered through the computation and testing of those meaningful, single-degree-of-freedom linear contrasts that were "built-in" to the experiment when the particular treatments were chosen by the experimenter. In many cases the set of linear contrasts will be orthogonal as well as meaningful. For examples of experiments for which the design and objectives suggest meaningful, perhaps orthogonal, single-degree-of-freedom linear contrasts to explain variation among treatments, see Bryan-Jones and Finney (1983), Carmer (1978), Chew (1976, 1977), Dawkins (1983), Johnson and Berger (1982), Little (1978, 1981), Mead and Pike (1975), Nelson and Rawlings (1983), or Petersen (1977).

There are, on the other hand, some experiments that the experimenter designs with the intent of examining the differences between members of each pair of treatments. Common examples of such a situation are performance trials to evaluate sets of crop cultivars. Other examples include herbicide, fungicide, insecticide, and other pesticide screening trials. Here pairwise comparisons may be sensible and meaningful, and it may well be a logical part of the experimental plan to perform them.

The purposes of this paper are: 1) to review and define the types and frequency rates of statistical errors with regard to pairwise multiple comparisons, 2) to describe a number of the pairwise multiple comparisons procedures that are available, and 3) to suggest that the least significant difference is always the procedure of choice when the appropriate contrasts among treatments each involve only two of the treatment means. We hope readers will find this presentation less confusing and more satisfactory than those given in statistical textbooks oridinarily used in teaching courses on the design and analysis of agronomic experiments.

TYPES AND RATES OF STATISTICAL ERRORS FOR PAIRWISE COMPARISONS

Let the true difference between two treatment means be represented by:

$$\delta_{ij} = \tau_i - \tau_j$$

where τ_i and τ_j represent the true effects of the *i*th and *j*th treatments, respectively. With the use of a pairwise multiple comparisons procedure one of three possible decisions is made concerning each pair of means; i.e., each δ_{ij}. The possible decisions are: 1) $\delta_{ij} < 0$; or 2) $\delta_{ij} = 0$; or 3) $\delta_{ij} > 0$.

The correctness of a particular decision based on a pair of observed means depends on the true or parameter values of the means. The latter are, in general, unknown. Several kinds of incorrect decisions or errors are possible (Table 1). If the parameter values of two means are really equal, i.e., $\delta_{ij} = 0$, reaching decision 1 or 3 on the basis of observed means results in a Type I error, which occurs when a true null hypothesis is rejected. On the other hand a Type II error occurs when a false null hypothesis is not rejected. Thus reaching decision 2 on the basis of observed means results in a Type II error if the two true means really are not equal, i.e., $\delta_{ij} \neq 0$. Still another kind of error is committed if decision 1 is reached, but decision 3 is actually correct, or if decision 3 is reached, but decision 1 is actually correct. These are called reverse decisions or Type III errors. In summary then, for any given pair of treatments, the experimenter will either make the correct decision or one of the three types of errors.

Table 1. Types of statistical errors possible when comparing two observed treatment means.

Decision based on observed means	True situation		
	$\delta_{ij} < 0$	$\delta_{ij} = 0$	$\delta_{ij} > 0$
1. $\delta_{ij} < 0$	Correct decision	Type I error	Type III error
2. $\delta_{ij} = 0$	Type II error	Correct decision	Type II error
3. $\delta_{ij} > 0$	Type III error	Type I error	Correct decision

[1] Contribution from the Dep. of Agronomy, Univ. of Illinois, 1102 S. Goodwin Ave., Urbana, IL 61801.

[2] Professor of biometry and professor of biometry and soil fertility, Dep. of Agronomy, Univ. of Illinois, Urbana.

When experimenters select a significance level, α, they are, in effect, stating the frequency of Type I errors that they are willing to accept. There are at least two ways of expressing this frequency or Type I error rate. The comparisonwise Type I error rate is defined as:

$$\alpha_C = \frac{\text{No. of Type I errors}}{\text{No. of comparisons for which the true difference} = 0}$$

while the experimentwise Type I error rate is defined as:

$$\alpha_E = \frac{\text{No. of experiments with one or more Type I errors}}{\text{No. of experiments with at least one true difference} = 0}.$$

When considering the experimentwise error rate, it is helpful to state the family of contrasts to which the experimentwise error rate is applicable. The definition given above applies to the family consisting of the set of all pairwise comparisons among the treatments. Other families are possible; for example, the family of interest might be a set of meaningful orthogonal linear contrasts among the means. In another situation, the family might be the set of contrasts which compares each treatment to a control or check or standard. In certain cases, the largest conceivable family might be appropriate; i.e. the family which consists of the set of all possible contrasts among the treatment means.

For a family of $q = (p - 1)$ orthogonal contrasts among p treatments the comparisonwise and experimentwise Type I error rates are related as follows:

$$\alpha_E = 1 - (1 - \alpha_C)^q;$$

$$\alpha_C = 1 - (1 - \alpha_E)^{1/q}.$$

Thus with $p = 15$ equal treatments, $\alpha_E = 0.5123$ if $\alpha_C = 0.05$. On the other hand $\alpha_C = 0.00365$ if $\alpha_E = 0.05$. The meaning is that, when each of 14 orthogonal contrasts are tested at the 5% significance level, the probability that at least one of the 14 contrasts will be incorrectly declared significant is 51.23%. If the researcher wished this probability to be only 5%, the test of each individual contrast would have to be performed at the 0.365% significance level.

In the case of pairwise multiple comparisons among p treatment means the family consists of the $m = p(p - 1)/2$ nonorthogonal contrasts between the members of pairs. The corresponding relationships between error rates are:

$$\alpha_E \le 1 - (1 - \alpha_C)^m < m\alpha_C;$$

$$\alpha_C \ge 1 - (1 - \alpha_E)^{1/m} > \alpha_E/m.$$

Thus with $p = 15$ equal treatments $m = 105$ and $\alpha_E \le 0.9954$ if $\alpha_C = 0.05$, and $\alpha_C > 0.000488$ if $\alpha_E = 0.05$. More exact values can be determined from tables provided by Harter (1957) and Harter et al. (1959). For example, if the $p = 15$ equal treatments are replicated four times in a randomized complete block design, α_E is equal to about 0.78 when $\alpha_C = 0.05$ and $\alpha_C = 0.000834$ when $\alpha_E = 0.05$.

The experimentwise Type I error rate is thought to be of considerable importance by some statisticians and researchers (e.g., Gill, 1973), but its use deprives the experimenter of the opportunity to select a comparisonwise significance level in accordance with his or her own assessment of the seriousness of Type I and III errors relative to the seriousness of Type II errors.

Comparisonwise and experimentwise expressions can be developed for both Type II and Type III error rates, but generally these are expressed only on a comparisonwise basis, if they are expressed at all. Thus the comparisonwise Type II error rate is:

$$\beta_C = \frac{\text{No. of Type II errors}}{\text{No. of comparisons for which the true difference} \neq 0}$$

and represents the comparisonwise probability of failure to reject a false hypothesis of no difference between two means. Similarly, the comparisonwise Type III error rate is:

$$\gamma_C = \frac{\text{No. of Type III errors}}{\text{No. of comparisons for which the true difference} \neq 0}$$

and represents the comparisonwise probability of concluding either that $\delta_{ij} > 0$ when in fact $\delta_{ij} < 0$, or that $\delta_{ij} < 0$ when in fact $\delta_{ij} > 0$.

The power of a statistical test of a null hypothesis is usually defined as the probability of rejecting the hypothesis when the hypothesis is false. Thus, for a pairwise comparison the comparisonwise power is equal to $[1 - \beta_C]$ if $\delta_{ij} \neq 0$. When the magnitude of δ_{ij} is very large the power approaches a value of 1.0; on the other hand, as the magnitude of δ_{ij} approaches 0.0, the power approaches the value of α_C. The comparisonwise correct decision rate for a pairwise comparison is equal to $[1 - \alpha_C]$ if $\delta_{ij} = 0$, and is equal to $[1 - \beta_C - \gamma_C]$ if $\delta_{ij} \neq 0$.

PAIRWISE MULTIPLE COMPARISON PROCEDURES

For some experiments sensible treatment contrasts are the pairwise comparisons among the observed means. As mentioned earlier, examples of such experiments include trials for evaluation of the performance of cultivars, or herbicides, fungicides, insecticides, or other pesticides. Unfortunately for researchers conducting such experiments, statisticians have not been able to reach general agreement as to which of several suggested procedures is the best for researchers to apply when pairwise multiple comparisons are appropriate.

Our Recommendation: The Least Significant Difference

Use of the least significant difference, LSD, dates back to R. A. Fisher and the early days of analysis of variance. The LSD procedure is really only a short-cut version of performing a series of t tests on all the possible pairs of treatment means.

A t test of the difference between the *i*th and *j*th observed treatment means can be computed as

$$t = (\overline{Y}_{i.} - \overline{Y}_{j.})/s_d$$

where s_d is the standard error of the difference between the two means. Instead of calculating the individual t values for all possible pairs of means and comparing each to the appropriate table value of t, the LSD is computed as that difference between two means which equals the table value of t multiplied by s_d. Thus the ordinary LSD, which is often referred to as the multiple t test, the unprotected LSD, or the unrestricted LSD, is computed as:

$$\text{LSD} = t\,(\alpha, f)\, s_d$$

where $t\,(\alpha, f)$ is the tabular value of "Student's t" for the selected significance level, α, and the degrees of freedom, f, associated with the standard error of the difference between two means, s_d. Observed means which differ by more than the LSD value are said to be significantly different.

Some statisticians have criticized the LSD procedure on the grounds that, if all the true treatment means are equal, the probability of falsely declaring significant the difference between the largest and smallest observed means (i.e., commission of a Type I error) is greater than α when the experiment includes more than two treatments. In other words the experimentwise Type I error rate associated with the LSD is greater than the comparisonwise Type I error rate which is the experimenter's selected significance level, α. If $p = 15$ equal treatments are replicated four times in a randomized complete block design the experimentwise error rate, α_E will be about 0.78 when the LSD is applied at a significance level with $\alpha = \alpha_C = 0.05$.

R. A. Fisher suggested that the experimenter could protect against high experimentwise error rates by performing a preliminary F test of treatment differences based on the ratio of the among treatments mean square divided by the error mean square. If the computed F value is declared significant, indicating real treatment effects, the experimenter computes the usual LSD; however, if the F value is not significant, indicating a lack of real differences among the treatment means, the experimenter makes no pairwise comparisons among the means, thus eliminating the possibility of making Type I errors in that particular experiment. This procedure is variously known as Fisher's least significant difference or FLSD, the protected LSD, or the restricted LSD. The critical difference is computed as:

$$\text{FLSD} = \text{LSD} = t(\alpha, f)\, s_d,$$

if the computed F ratio is significant, or

$$\text{FLSD} = \infty,$$

if the computed F ratio is not significant.

The requirement of a significant F ratio results in the reduction of the comparisonwise Type I error rate to a value less than the stated significance level of the LSD. Empirical demonstrations of this are provided by Carmer and Swanson (1971, 1973) and Bernhardson (1975). A formalized mathematical proof is given by Smith and Han (1981).

In deciding whether to use the ordinary LSD or the restricted LSD, the experimenter needs to consider the question: "How likely is it that all p treatments in my experiment have exactly the same true means?" If it is quite unlikely that all p treatment means are equal, there may be little or no point in requiring the analysis of variance F ratio to be significant. On the other hand, if the experimenter has evidence that all p treatment means might be expected to be equal, use of the restricted LSD may be a good choice.

Consider two examples. In one case, a forage breeder wishes to compare yields of eight genetically similar alfalfa (*Medicago sativa* L.) clones. Since they are genetically similar and the overall hypothesis tested by the analysis of variance F ratio might be true, the breeder decides to use the restricted LSD. The second example concerns a performance trial in which there are 250 commercial corn (*Zea mays* L.) hybrids produced by 18 different seed companies. In this case a hybrid produced by one company might be genetically identical to a hybrid produced by another company, but it is impossible to imagine that all 250 hybrids are identical. Thus the overall hypothesis tested by the analysis of variance F ratio is known to be false before the trial even starts. For an experiment like this, a nonsignificant F value is more apt to result from poor precision due to poor design and/or poor conduct of the experiment than from a true null hypothesis. If so, the experimenter should be much more concerned about improving the precision of the experiment than about whether to use the restricted LSD rather than the ordinary LSD.

The least significant difference is based on a comparisonwise Type I error rate and its use is justified when the individual comparisons within pairs of treatments are the conceptual units of interest to the experimenter. Carmer and Walker (1982) described a situation where the experimenter wished to compare each of 15 cultivars to each of the other 14; i.e., 105 pairwise comparisons were to be made. The experimenter conducted 105 trials with each trial consisting of four replications of one pair of cultivars. Data from each trial were subjected to analysis of variance, and the LSD for comparing the two means was calculated at the 5% significance level. A statistician later advised the experimenter to repeat the study using four replications of a randomized complete block design with 15 treatments. The experimenter was subsequently criticized by peer scientists for using the LSD for this second experiment because it resulted in an experimentwise error rate, α_E, of about 0.78. However, as Carmer and Walker (1982) point out, the experimentwise error rate for the first study involving 105 independent trials had an expected value of

$$\alpha_E = 0.9954 = [1 - (0.95)^{105}].$$

Based on this fact, it is quite clear that if the individual comparison, and not the experiment, is the conceptual unit of interest, then the experimenter should not be penalized for using an efficient experimental design; the penalties imposed by use of an experimentwise error rate should not be inflicted upon the experimenter because he/she used a design with 60 experimental units (15 cultivars × 4 replicates) rather than 105 trials occupying 840 experimental units. Duncan and Brant (1983) state: "In the simultaneous testing of m comparisons in one experiment, the objectives are the same as if each test were being made in a separate experiment." O'Brien (1983) has also written in favor of comparisonwise error rates. The penalties of using experimentwise Type I error rates include larger critical values than the LSD with a resultant larger Type II error rate, smaller power, and smaller correct decision rate when $\delta_{ij} \neq 0$. Thus in addition to its simplicity of calculation the LSD is a more powerful technique that is more sensitive to treatment effects.

Through a simulation study Carmer (1976) demonstrated that the choice of significance level for the LSD does directly influence the comparisonwise Type II and Type III error rates for true differences of given magnitudes. Analytically these rates may be expressed as:

$$\beta_C = \text{Probability}$$
$$\{-[t(\alpha_C,f) + \delta_{ij}/\sigma_d] < t < [t(\alpha_C,f) - \delta_{ij}/\sigma_d]\}$$

and

$$\gamma_C = \text{probability } [t > t(\alpha_C,f) + \delta_{ij}/\sigma_d],$$

where σ_d is the true standard error of the difference between the two treatment means.

For the specific case of corn hybrid performance trials in Illinois Carmer (1976) provided an assessment of the seriousness of Type I and Type III errors relative to Type II errors; through minimization of the weighted average losses due to all three kinds of errors the optimal significance level was found to be in the range of $\alpha_C = 0.20$ to 0.40. For other kinds of experiments researchers need to consider the relative seriousness of the several types of statistical errors and then use the results of such assessments in selecting significance levels which minimize losses due to such errors.

Regardless of the specific value of $\alpha_C > 0$ which is selected, the probability of committing at least one Type I error will increase as the number of comparisons for which $\delta_{ij} = 0$ increases. Put another way, this simply means that the more decisions a decision-maker makes, the more likely at least one decision will be wrong. To us, this idea seems only reasonable.

Other Procedures: Some Worse than Others

The invention of experimentwise Type I error rate by theoretical statisticians has spurred seemingly never-ending research by a generation or more of mathematical statisticians in an attempt to find a procedure with more attractive properties than the LSD. This effort has resulted in a large number of procedures, some of which are useful if the experiment rather than the individual comparison is the conceptual unit of interest to the experimenter.

Family of All Pairwise Comparisons: Tukey's Test

In response to the previously mentioned criticism of the high experimentwise error rate associated with the ordinary LSD, J. W. Tukey developed a procedure based on the researcher's selection of an experimentwise Type I error rate. This procedure is known as Tukey's w procedure, Tukey's Significant Difference or TSD, and the Honestly Significant Difference or HSD. The critical difference is computed as:

$$w = \text{TSD} = \text{HSD} = Q(\alpha, p, f)\, s_d/\sqrt{2}$$

where Q (α, p, f) is the appropriate value from a table of studentized ranges for the selected significance level, α, with p treatments and f degrees of freedom associated with the estimate of experimental error.

For $p > 2$ the critical value of Tukey's procedure is larger than the LSD, but for $p = 2$:

$$\begin{aligned} \text{LSD} &= w = \text{TSD} = \text{HSD} \\ &= Q(\alpha, p, f)\, s_d/\sqrt{2} = t(\alpha, f)\, s_d. \end{aligned}$$

Performance of Tukey's procedure at $\alpha = 0.05$ results in, on the average, the commission of at least one Type I error in 5 out of 100 experiments when the p true treatment means are all equal. With $p = 15$ equal treatments replicated four times in a randomized complete block design the comparisonwise error rate will be $\alpha_C = 0.000834$ when Tukey's procedure is applied at a significance level of $\alpha = \alpha_E = 0.05$. That is, Tukey's test at the 5% level gives the same result as if the LSD was performed at the 0.0834% level.

The experimentwise error rate employed with Tukey's procedure applies to the family of all pairwise comparisons among the treatment means.

Family of Comparisons of All Treatments with a Control: Dunnett's Test

In some experiments the family of comparisons of interest is not the set of all possible pairwise comparisons. In Florida sugarcane (*Saccharum* spp.) performance trials, for example, it is common to compare each cultivar being evaluated to a standard or check cultivar rather than to each other. Some statisticians argue that in such a case there is need for a procedure based upon an experimentwise Type I error rate for the family of $q = (p - 1)$ comparisons between each cultivar and the check. Dunnett (1955, 1964) has devised a procedure to do this and prepared special tables of Dunnett's t for experimentwise Type I error rates of $\alpha_E = 0.01$ and 0.05. The critical value for Dunnett's test is computed as:

$$\text{DSD} = t\,(\text{Dunnett}, \alpha, q, f)\, s_d$$

where t (Dunnett, α, q, f) is the appropriate tabulated value for the selected significance level, α, with $q = (p - 1)$ treatments excluding the standard and f degrees of freedom associated with the estimate of experimental error.

Performance of Dunnett's procedure at $\alpha = 0.05$ results in, on the average, the commission of at least one Type I error in 5 out of 100 experiments when the $q = (p - 1)$ treatments are all equal to the check.

With $q = 15$ treatments all equal to the standard and replicated four times in a randomized complete block design the comparisonwise error rate will be $\alpha_C = 0.00503$ when Dunnett's procedure is applied at a significance level of $\alpha = \alpha_E = 0.05$. That is, Dunnett's test at the 5% level gives the same result as if the LSD was performed at the 0.503% level.

For those experimenters for whom the conceptual unit of interest is the experiment Dunnett's procedure may be acceptable. However, as with the case of all possible pairwise comparisons, if the individual comparison is the unit of interest, then the LSD should be the procedure of choice. Under no circumstances should Dunnett's test be used for making all possible pairwise comparisons.

Family of All Possible Contrasts: Scheffe's Test

The significance level selected for Scheffe's method is also an experimentwise error rate, but it is based upon a much different family of comparisons than either Tukey's or Dunnett's procedure. Scheffe's method applies to the family of all possible linear contrasts among the treatment means which, clearly, is a much larger family than the subset of all possible pairwise comparisons. Consequently, the critical value for Scheffe's test is even larger than that for Tukey's procedure. The method is quite general in that it is applicable to any imaginable linear contrast among treatment means, and is sometimes used as a pairwise multiple comparisons procedure (even though such usage is not recommended). However, when employed in this latter context, the critical value is:

$$\text{SSD} = [q, F(\alpha, q, f)]^{1/2}\, s_d$$

where F (α, q, f) is the tabular F value with $q = (p - 1)$ and f degrees of freedom at the α significance level. Note that for $p = 2$, the SSD and LSD are equal, but for $p > 2$ the relationship SSD > TSD > LSD holds.

For $p = 15$ equal treatments replicated four times in a randomized complete block design the comparisonwise error rate is extremely small; $\alpha_C = 0.00000546$ when Scheffe's procedure is applied at a significance level of $\alpha = \alpha_E = 0.05$. That is, Scheffe's test at the 5% level gives the same result as if the LSD was performed at the 0.000546% level.

While Scheffe's procedure is probably the least-suited method for pairwise multiple comparisons, or for other linear contrasts specified prior to the conduct of the experiment, its conservative nature and basis of dealing with the family of all imaginable linear contrasts give the procedure some appeal for those situations where the investigator is involved in "data snooping," exploratory data analysis, or looking at contrasts suggested by the data; in these cases the appropriate standard error associated with the contrast of concern would be substituted for s_d in the calculations. While the use of $\alpha_C = 0.05$ may be unacceptable in these cases of data selection (as opposed to a priori contrasts), it is arguable whether a comparisonwise error rate as low as that associated with Scheffe's procedure is either necessary or desirable.

The Waller-Duncan Bayesian k-ratio t Test: The Bayes Least Significant Difference

Efforts by Duncan and his students subsequent to his development of Duncan's multiple range test have helped to clarify the multiple comparisons problem. Their approach requests the experimenter to give some thought to the "costs" associated with the making of the several types of statistical errors and requires the experimenter to select a value of k which is a ratio that quantifies the seriousness of Type I and III errors relative to Type II errors. Duncan (1965) asserted that values of k-ratios equal to 500, 100, and 50 are roughly comparable to significance levels of $\alpha = 0.01$, 0.05, and 0.10, respectively. Carmer (1976) indicated that k-ratios equal to 20, 7, and 2 correspond to significance levels of $\alpha = 0.20$, 0.40, and 0.80, respectively. The theoretical development of the procedure by Waller and Duncan (1974) includes use of Bayesian statistical principles in the examination of prior probabilities of decision errors. Because of its Bayesian properties and its similarities to the least significant difference, the procedure is often known as the Bayes least significant difference. The single-valued critical difference is the product of the standard error of the difference between two treatment means and a quantity called the minimum average risk t value. This t value differs from Student's t, but depends, instead, upon the magnitude of the analysis of variance F value calculated to test overall treatment effects and upon the particular k-ratio selected as well as upon the treatment and error degrees of freedom. For pairwise comparisons the critical value for the Bayes least significant difference is:

$$\mathrm{BLSD} = t\,(k, F, f, q)\, s_d$$

where t (k, F, f, q) is the tabulated minimum average risk t value for the selected k-ratio, the numerical value of the computed F ratio, and the degrees of freedom, f and q, for error and treatments, respectively.

Tables of minimum average risk t values for k-ratios equal to 500, 100, and 50 are given in Waller and Duncan (1969); these tables turned out to contain errors and corrected tables appeared later (Waller and Duncan, 1972). Waller and Kemp (1975) described a computer program which computes the minimum average risk t value for a specified combination of k, F, f, and q.

An application of the Bayes least significant difference to means from an agronomic experiment by Smith (1978) illustrates the effect that the magnitude of the analysis of variance F value has on the BLSD. In his example the FLSD = 436 and the HSD = 639; with a F value of 4.43 the BLSD = 445 and is comparable to the FLSD. However, with a F value of 1.4, the BLSD = 600 and is nearly equal to the HSD. This example demonstrates the property of the BLSD which allows it to avoid Type I errors when the F value is small and to avoid Type II errors when the F value is large.

Table 2. Effects of the magnitude of the computed F value on the minimum average risk t value and comparisonwise error rate.

Computed F value	Minimum avg. risk t value	Comparisonwise error rate, α_C
0.968	3.52	0.0011
1.935	2.66	0.0110
3.870	2.12	0.0400
5.805	1.99	0.0531
7.744	1.93	0.0604
9.675	1.90	0.0643
19.350	1.84	0.0728

Perhaps the most essential difference between the ordinary LSD and the BLSD is that with the former procedure the experimenter selects the significance level at which to test treatment differences, while with the BLSD the observed data are allowed to have a large influence on the comparisonwise error rate. For example, with p = 15 equal treatments replicated four times in a randomized complete block design, the ordinary LSD will have $\alpha_C = 0.05$ when the LSD is performed at the 5% significance level. But for the BLSD with a k-ratio of 100, the value of α_C will depend on the value of the minimum average risk t value which, in turn, depends on the magnitude of the analysis of variance F value, as is shown in Table 2. For the FLSD performed at the 5% significance level, the value of the comparisonwise error rate is 0.05 for all computed F values greater than the appropriate tabulated F value, (1.935 in this case), but for non-significant F values the value of $\alpha_C = 0.00$. For the BLSD, however, the value of α_C increases as the computed F value increases. Note that for F = 1.935 the comparisonwise error rate would be $\alpha_C = 0.05$ for the FLSD but $\alpha_C = 0.0110$ for the BLSD.

In our opinion the primary value in the development of the BLSD by Duncan and his colleagues has been to draw the researcher's attention to the importance of assessing the relative seriousness of the various types of statistical errors. As was mentioned earlier in this paper, Carmer (1976) has shown that such assessments can be used to determine an optimal significance level for the LSD. With the LSD the researcher has the opportunity to assess the relative seriousness of statistical errors and to manage the risks associated with them through his or her selection of a significance level. With the BLSD, on the other hand, through selection of the k-ratio the researcher only assesses the risks of statistical errors; the data, rather than the researcher, determine the choice of significance level for management of the risks.

Other Procedures: Multiple Range Tests

The Student-Newman-Kuels and the Duncan multiple range tests operate in such a manner that, for the family of all pairwise comparisons, the experimentwise Type I error rate is intermediate between that obtained with the LSD and that obtained with the TSD.

While the LSD and TSD each require the calculation of a single critical value, the Student-Newman-Kuels procedure involves the computation of (p − 1) critical values:

$$\mathrm{SNK}_i = Q\,(\alpha, i, f)\, s_d/\sqrt{2}$$

for i = 2, 3, . . ., p and Q (α, i, f) is the appropriate value from a table of studentized ranges. The value of SNK_2 equals the LSD value while SKN_p equals the TSD value; for intermediate values of i, SNK_i is intermediate to the LSD and TSD.

The SNK, like the TSD, utilizes ordinary studentized ranges which are tabulated, for example, in Table II.2 of Harter et al. (1959). For Duncan's multiple range test, however, special studentized ranges with $\alpha_i = [1 - (1 - \alpha)^{i-1}]$ for i = 2, 3, ..., p, are employed.

Tables of the special studentized ranges were first presented by Duncan (1955) and later recomputed by Harter et al. (1959). For Duncan's multiple range test the (p − 1) critical values are calculated as:

$$\mathrm{DMRT}_i = Q(\alpha_i, i\,f)\,s_d/\sqrt{2}$$

for i = 2, 3, . . ., p. Except for i = 2, values of DMRT_i are larger than the LSD, but smaller than the TSD, and smaller than corresponding SNK_i values.

The present authors recommend that neither DMRT, nor SKN, nor any other multiple range procedures ever be used for comparisons among treatment means. We have several reasons. Perhaps the most important is that if the experimenter firmly believes that the experiment is the conceptual unit of interest, then appropriate procedures are Tukey's test for pairwise comparisons, Dunnett's test for comparisons of individual treatments with the control, and Scheffe's test for contrasts suggested by looking at the data. On the other hand, if the experimenter considers the individual comparisons to be the conceptual units of interest, then the LSD is the appropriate procedure. The multiple range tests require multi-valued critical values, which complicate the presentation of results, and are less powerful and less able to detect differences than the LSD. With multiple range tests the difference between two treatments required for significance depends on, p, the number of treatments in the experiment. As Carmer and Walker (1982) state, it does not make much sense to think that the true difference between two treatments depends in any way on what other treatments are included in the experiment. Even Duncan (1970) has stated that there are other procedures with more logical foundations than his 1955 multiple range test, DMRT. One other disadvantage of multiple range tests is that their use is not appropriate in the construction of confidence intervals; on the other hand procedures which employ only a single critical value are easily used in the construction of confidence intervals.

Still Other Procedures

Kirk (1982) discusses several other procedures for testing nonorthogonal contrasts among treatment means. One of these is referred to as Dunn's multiple comparisons procedure or the Bonferroni t procedure. For m contrasts the experimentwise Type I error rate cannot exceed the sum of the m comparisonwise error rates; i.e.:

$$\alpha_E \leq \Sigma\, \alpha_{C_i}$$

for i = 1, 2, . . ., m. Therefore, if each of the m contrasts is tested at the $\alpha_C = \alpha/m$ level of significance, the experimentwise Type I error rate cannot exceed α. For pairwise comparisons, the critical value for Dunn's procedure is:

$$\mathrm{LSD} = t(\alpha/m, f)\,s_d$$

where t (α/m, f) is the absolute value of Student's t with f degrees of freedom which will be exceeded due to chance with probability α/m. Since t values for $\alpha_C = \alpha/m$ will not generally be found in tabulations of Student's t for $\alpha = 0.01$ or 0.05, special tables have been prepared (Dunn, 1961, and Dayton and Schafer, 1973).

A modification to Dunn's procedure has been proposed by Sidak (1967) who utilized the property that the experimentwise error rate is never greater than $[1 - (1 - \alpha_C)^m]$. Instead of testing each contrast at $\alpha_C = \alpha/m$, Sidak's modification tests each contrast at $\alpha_C = [1 - (1 - \alpha)^{1/m}]$. Kirk (1982) calls this modification the Dunn-Sidak procedure; for pairwise comparisons the critical value is:

$$\mathrm{LSD} = t(1 - [1 - \alpha]^{1/m}, f)\,s_d$$

where $t(1 - [1 - \alpha]^{1/m}, f)$ is the absolute value of Student's t with f degrees of freedom which will be exceeded due to chance with probability $[1 - (1 - \alpha)^{1/m}]$. Again, if $\alpha = 0.01$ or 0.05, the value of $\alpha_C = [1 - (1 - \alpha)^{1/m}]$ will not generally be found in tabulations of Student's t. A table of critical values has been presented by Games (1977). The critical values of t for the Dunn-Sidak procedure are always smaller than those for the Dunn procedure; tables for both procedures are given in Kirk (1982).

The use of Dunn's procedure is equivalent to performing the ordinary LSD at the $\alpha_C = \alpha/m$ level of significance, while the Dunn-Sidak procedure is equivalent to performing the ordinary LSD at the $\alpha_C = [I - (1 - \alpha)^{1/m})$ significance level.

According to Kirk (1982), neither procedure is recommended for pairwise comparisons, but, if an experimenter considers the experiment to be the conceptual unit of interest, either procedure would be appropriate for a specific set of planned, but non-orthogonal, contrasts among means. On the other hand, if the individual contrasts are the conceptual units of interest, the use of the usual Student's t tests or the equivalent single degree of freedom F tests is appropriate.

Another Approach: Cluster Analysis of Treatment Means

An alternative to the use of pairwise multiple comparisons procedures is a technique known as cluster analysis. For some researchers cluster analysis is attractive because, unlike pairwise multiple comparisons, it results in non-overlapping, distinct, mutually exclusive groupings of the observed treatment means. One method of cluster analysis which uses a divisive criterion for subdividing a set of means into groups was proposed by Scott and Knott (1974); use of the Scott-Knott procedure in agricultural research has been suggested by Chew (1977), Gates and Bilbro (1978), and Madden et al. (1982). Illustrative examples are provided in each of these papers.

Willavize et al. (1980) and Carmer and Lin (1983) have expressed a need for caution in applying the Scott-Knott or other clustering criteria to experiments where use of the LSD is appropriate. In both papers the concerns raised are based on the results of simulation studies. In a comparison of the Scott-Knott and three agglomerative clustering criteria, Willavize et al. (1980) found that Type I error rates with the clustering procedures were often considerably higher than the stated significance level and often appreciably higher than with the restricted LSD. On the other hand, the clustering techniques produced higher correct decision rates than did the LSD for small true relative differences between means ($\delta_{ij}/\sigma \leq 1.5$), while all procedures performed quite well for large differences ($\delta_{ij}/\sigma > 2.2$). Carmer and Lin (1983) compared Type I error rates for the Scott-Knott and three divisive criteria based on F tests. Again, cluster analysis Type I error rates were found to be appreciably higher than with the LSD.

Another disconcerting finding about the clustering procedures is that, in both studies, evidence was presented which indicates that the Type I error rate associated with a particular clustering method is determined more by the precision of the experiment than by the selected significance level. Willavize et al. (1980) state that a cluster analysis procedure may be applicable in either of two cases: 1) when a significance level in the range $\alpha = 0.20$ to 0.40 is justified on the basis that Type I errors produce little harm, or 2) when the experiment has been performed with great precision. Thus the application by Madden et al. (1982) of the Scott-Knott method to maize dwarf mosaic virus susceptibility data from ten dent corn inbreds may be a much more appropriate use of the technique than applying it at the traditional 5% significance level to group corn hybrids on the basis of their yield performance.

A final comment about cluster analysis methods is that the computations are considerably more numerous and complex than for the single critical value of the LSD.

UNEQUAL REPLICATIONS AND/OR UNEQUAL VARIANCES

For a randomized complete block design the standard error of the difference between two treatment means is calculated as the square root of two times the error mean square divided by the number of replications; i.e.:

$$s_d = [2s^2/r]^{1/2}$$

where s^2 is the estimated error variance and r is the number of replications. If the *i*th and *j*th treatments have different numbers of replications, the standard error of the difference between the *i*th and *j*th means is:

$$s_{d(ij)} = [s^2/r_i + s^2/r_j]^{1/2}$$

and can be readily utilized in the formulae for either the LSD, TSD, or BLSD.

If the *i*th and *j*th treatments have unequal variances or both the variances and the numbers of replications are unequal, the standard error of the difference between the *i*th and *j*th treatment means becomes:

$$s_{d(ij)} = [s_i^2/r_i + s_j^2/r_j]^{1/2}$$

with f′ degrees of freedom, where

$$f' = [s_i^2/r_i + s_j^2/r_j]^2/[(s_i^4/r_i^2f_i) + (s_j^4/r_j^2f_j)]$$

where f_i and f_j are the degrees of freedom associated with s_i^2 and s_j^2, respectively. The calculation of f′ is based on Satterthwaite's (1946) approximation to the degrees of freedom associated with a linear combination of estimates of variance components. The pairwise comparison between the *i*th and *j*th treatment means then has critical differences as follows for the LSD, TSD, and BLSD procedures:

$$\mathrm{LSD(ij)} = t\,(\alpha, f')\, s_{d(ij)}$$

$$\mathrm{TSD(ij)} = Q\,(\alpha, p, f')\, s_{d(ij)}$$

$$\mathrm{BLSD(ij)} = t\,(k, F, f', q)\, s_{d(ij)}.$$

CONCLUDING REMARKS

Ten multiple comparisons procedures have been described in preceding sections. Of these, the least significant difference is the most appropriate for any comparison involving only two treatment means. Such comparisons are logical when there is no structure in the set of treatments (e.g., when the treatments are cultivars, or herbicides or other pesticides). The least significant difference is also the most appropriate for comparisons between individual treatments and a control or standard.

Most experiments, however, involve treatments which do have structure (e.g., levels of quantitative controlled factors such as rates of fertilizer, row spacings, plant densities, or dates or application). For these experiments the use of meaningful, single degree of freedom linear contrasts provides a statistical tool more powerful than any pairwise multiple comparisons procedure, including the least significant difference. Meaningful linear contrasts are more powerful (i.e., better able to detect treatment effects), because they usually involve a linear combination of more than two treatment means, while only two treatment means are included in a pairwise comparison.

Finally, a word of caution may be helpful in regard to experiments involving several rates of each pesticide or other chemical material being evaluated in a trial. Regression analysis of trends will be more informative and statistically more sound for evaluation of the response to differing rates of a particular pesticide or other chemical material. A pairwise multiple comparisons procedure is not the best statistical technique for interpreting the variation in response due to differing rates of application. Pairwise comparisons between two materials applied at equivalent rates may be meaningful, but for differing rates within a material the response trend is of main interest and concern.

REFERENCES

1. Bernhardson, C. S. 1975. Type I error rates when multiple comparison procedures follow a significant F test ANOVA. Biometrics 31:229-232.
2. Bryan-Jones, J., and D. J. Finney. 1983. On an error in "Instructions to authors". Hortscience 18:279-282.
3. Carmer, S. G. 1976. Optimal significance levels for application of the least significant difference in crop performance trials. Crop Sci. 16:95-99.
4. ----. 1978. Use of MATRIX for single degree of freedom contrasts in factorial experiments. p. 284-287. *In* R. H. Strand (ed.) Proceedings of the Third Annual Conference of the SAS Users Group International. SAS Institute Inc., Raleigh, NC.
5. ----, and W. T. Lin. 1983. Type I error rates for divisive clustering methods for grouping means in analysis of variance. Commun. Stat.-Simul. Comput. B12:451-466.
6. ----, and M. R. Swanson. 1971. Detection of differences between means: A Monte Carlo study of five multiple comparison procedures. Agron. J. 63:940-945.
7. ----, and ----. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J. Am. Stat. Assoc. 68:66-74.
8. ----, and W. M. Walker. 1982. Baby Bear's dilemma: A statistical tale. Agron. J. 74:122-124.
9. Chew, V. 1976. Comparing treatment means: A compendium. Hortscience 11:348-357.
10. ----. 1977. Comparisons among treatment means in an analysis of variance. USDA Rep. ARS/H/6.
11. Dawkins, H. C. 1983. Multiple comparisons misused: Why so frequently in response-curve studies. Biometrics 39:789-790.
12. Dayton, C. M., and W. D. Schafer. 1973. Extended tables of t and chi-square for Bonferroni tests with unequal error allocation. J. Am. Stat. Assoc. 68:78-83.

13. Duncan, D. B. 1955. Multiple-range and multiple-F tests. Biometrics 11:1-42.
14. ----. 1965. A Bayesian approach to multiple comparisons. Technometrics 7:171-222.
15. ----. 1970. Multiple comparison methods for comparing regression coefficients. Biometrics 26:141-143.
16. ----, and L. J. Brant. 1983. Adaptive t tests for multiple comparisons. Biometrics 39:790-794.
17. Dunn, O. J. 1961. Multiple comparisons among means. J. Am. Stat. Assoc. 56:52-64.
18. Dunnett, C. W. 1955. A multiple comparisons procedure for comparing several treatments with a control. J. Am. Stat. Assoc. 50:1096-1121.
19. ----. 1964. New tables for multiple comparisons with a control. Biometrics 20:482-491.
20. Gates, C. E., and J. D. Bilbro. 1978. Illustration of a cluster analysis method for mean separation. Agron. J. 70:462-465.
21. Games, P. A. 1977. An improved table for simultaneous control on g contrasts. J. Am. Stat. Assoc. 72:531-534.
22. Gill, J. L. 1973. Current status of multiple comparisons of means in designed experiments. J. Dairy Sci. 56:973-977.
23. Harter, H. L. 1957. Error rates and sample sizes for range tests in multiple comparisons. Biometrics 13:511-536.
24. ----, D. S. Clemm, and E. H. Guthrie. 1959. The probability integrals of the range and of the studentized range—Probability integral and percentage points of the studentized range; critical values for Duncan's new multiple range test. Wright-Patterson Air Force Base: Wright Air Development Center Technical Report 58-484, vol II.
25. Johnson, S. B., and R. D. Berger. 1982. On the status of statistics in Phytopathology. Phytopathology 72:1014-1015.
26. Kirk, R. E. 1982. Experimental design: Procedures for the behavioral sciences. 2nd ed. Brooks/Cole Publishing Co., Monterey, CA.
27. Little, T. M. 1978. If Galileo published in Hortscience. Hortscience 13:504-506.
28. ----. 1981. Interpretation and presentation of results. Hortscience 16:637-640.
29. Madden, L. V., J. K. Knoke, and R. Louie. 1982. Considerations for the use of multiple comparison procedures in phytopathological investigations. Phytopathology 72:1015-1017.
30. Mead, R., and D. J. Pike. 1975. A review of response surface methodology from a biometrics viewpoint. Biometrics 31:803-851.
31. Nelson, L. A., and J. O. Rawlings. 1983. Ten common misuses of statistics in agronomic research. J. Agron. Educ. 12:100-105.
32. O'Brien, P. C. 1983. The appropriateness of analysis of variance and multiple comparison procedures. Biometrics 39:787-788.
33. Petersen, R. G. 1977. Use and misuse of multiple comparison procedures. Agron. J. 69:205-208.
34. Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biometrics Bull. 2:110-114.
35. Scott, A. J., and M. Knott. 1974. A cluster analysis method for grouping means in analysis of variance. Biometrics 30:507-512.
36. Sidak, Z. 1967. Rectangular confidence regions for the means of multi-variate normal distributions. J. Am. Stat. Assoc. 62:626-633.
37. Smith, C. W. 1978. Bayes least significant difference: A review and comparison. Agron. J. 70:123-127.
38. Smith, W. C., and C. P. Han. 1981. Error rate for testing a contrast after a significant F test. Commun. Stat.-Simul. Computa. B10:545-556.
39. Waller, R. A., and D. B. Duncan. 1969. A Bayes rule for the symmetric multiple comparisons problem. J. Am. Stat. Assoc. 64:1484-1503.
40. ----, and ----. 1972. Corrigenda. J. Am. Stat. Assoc. 67:253-255.
41. ----, and ----. 1974. A Bayes rule for the symmetric multiple comparisons problem II. Ann. Instit. Stat. Math. 26:247-264.
42. ----, and K. E. Kemp. 1975. Computations of Bayesian t-values for multiple comparisons. J. Stat. Comput. Simul. 4:169-171.
43. Willavize, S. A., S. G. Carmer, and W. M. Walker. 1980. Evaluation of cluster analysis for comparing treatment means. Agron. J. 72:317-320.