Letter to the Editor

Considerations for the Use of Multiple Comparison Procedures in Phytopathological Investigations

L. V. Madden, J. K. Knoke, and Raymond Louie

Systems specialist, Department of Plant Pathology; research entomologist and research plant pathologist, U.S. Department of Agriculture, Agricultural Research Service, respectively, all at the Ohio Agricultural Research and Development Center, Wooster 44691.
Cooperative investigations of the Ohio Agricultural Research and Development Center (OARDC) and the U.S. Department of Agriculture, Agricultural Research Service. Approved for publication as Journal Article 14-82 of OARDC.

This paper reports the results of research only. Mention of a commercial or proprietary product does not constitute an endorsement by the USDA.

We thank J. J. Abt, R. J. Anderson, and S. S. Mendiola for technical assistance.

Accepted for publication 30 March 1982.

Plant pathologists commonly summarize data by analysis of variance (ANOVA) and multiple comparison (MC) of means. Presentation of results in PHYTOPATHOLOGY typically consists of a table of treatment means followed by letters; those followed by the same letter are not significantly different. Using these MC procedures may be inappropriate in many situations (4,7,8,10,11,13). Petersen (13) stated "For experiments involving factorial sets of treatments or graded levels of quantitative factors, there is almost always a statistical procedure which can be specified in advance and which is more appropriate than a multiple comparison test." Petersen (13) then estimated that in 70% of the papers in Volume 67 of the *Agronomy Journal* in which MC procedures were used, the procedures were either entirely inappropriate or not the most meaningful way of analyzing the data. Recently, Johnson and Berger (9) stated that MC procedures were used inappropriately in two-thirds of the tables and figures in PHYTOPATHOLOGY. This degree of misuse probably is typical among most agricultural journals (4,7,10). Petersen (13) and others (4,7,10) expressed the view of many statisticians that designed comparisons, preferably orthogonal contrasts, should be used for determining treatment effects. Nevertheless, these authors concede that MC tests are useful for grouping means from experiments involving unstructured, qualitative treatments, eg, cultivars.

Numerous procedures have been suggested for MCs of means. These procedures include: least significant difference (LSD); Fisher's least significant difference (FLSD), ie, only comparing means if the *F*-test from ANOVA is significant; Tukey's significant difference (TSD); Student-Newman-Keuls (SNK); Scheffe's significant difference (SSD), which is a special form of a method for testing any linear contrast; Duncan's multiple range test (DMRT), which may be referred to as Duncan's new multiple range test; Duncan's least significant difference (DLSD); and the Waller-Duncan Bayesian least significant difference (BLSD) (2–5). Reviews and comparisons of the statistical properties of these MC procedures have been published (3,4,12). In any MC procedure, if the observed difference between any two means is greater than a *critical value*, the two means are considered to be different. In general, all possible differences of the means are calculated, and the significant differences determined. The magnitudes of the critical values for each of these procedures vary and therefore the results also vary!

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. § 1734 solely to indicate this fact.

The purpose of this article is to demonstrate the different results obtained by using these MC procedures, to discuss some of the properties of these procedures, and to present an alternative to MCs for analyzing qualitative treatments.

Example. Ten dent corn (*Zea mays* L.) inbreds were evaluated for susceptibility to maize dwarf mosaic virus (MDMV) at Wooster, OH, during 1981. Two 25-plant rows were planted with each cultivar in each of three replications and mechanically inoculated with MDMV (strain B) by using an artist's airbrush. The proportion of plants infected by MDMV 63 days after planting was assessed in each row. Analysis of the data revealed that the traditional arcsin, square-root transformations were not necessary because there were no statistically significant relationships between cultivar means and their variances.

Multiple comparisons of the means of the inoculated rows are presented in Table 1. Since differences among cultivar means were statistically significant and the LSD and FLSD results were the same, only FLSD results are shown. The critical values needed to declare means significantly different ranged from 0.064 to 0.156. The set of critical values used in the SNK and DMRT procedures depend on the range of the means being tested; therefore only the critical value extremes are shown. The number of statistically significant difference pairs ranged from 19 to 30. With 10 means, there is a total of 45 difference pairs. Identical mean separations in this example were obtained for DLSD and BLSD and also for FLSD and DMRT. All of the procedures found cultivar 10 different from the rest. Several of the procedures found that the means for cultivars 7, 8, and 9 were not different from each other, but that they were significantly different from the rest. The separation of cultivars 2 through 6 varied considerably among the MC procedures.

Multiple comparisons. The choice of MC procedure depends on the beliefs (biases) of the statistician or, more likely, tradition of the professional society. Perusal of PHYTOPATHOLOGY will demonstrate that most plant pathologists use DMRT, whereas usage of the TSD, SSD, and SNK procedures is almost nonexistent. For example, DMRT was used in 65% of the papers in Vol. 70 of PHYTOPATHOLOGY in which MC procedures were used. Another 18% of the papers with MCs did not mention the procedure used. Each procedure has advantages and disadvantages; statisticians have not yet developed an ideal method that fits all situations.

To compare these MC procedures, an understanding of the types of errors that can be committed is required. If the true means for treatments j and k are μ_j and μ_k, the three possible decisions are: (i) $\mu_j = \mu_k$; (ii) $\mu_j > \mu_k$; and (iii) $\mu_j < \mu_k$. The true means are estimated from the data and tested with an MC procedure to make one of the decisions. If the true values of the means are equal, reaching decision ii or iii is called a type I error. If the means are not equal, reaching decision i is called a type II error. A type III error is committed if decision ii is reached when iii is true, or if decision iii is reached when ii is true. Most of the MC procedures were developed to control the rate of type I errors for the collection of comparisons within an experiment, ie, the experimentwise type I error rate. Only the LSD procedure contains no provision for controlling this error. The BLSD and DLSD procedures attempt to control both type I and II error rates by using Bayesian statistical theory to evaluate prior probabilities of decision errors (5,16).

Carmer and Swanson (2,3) determined the error rates for all of these procedures through extensive simulation (Monte Carlo) studies. They found the SSD, TSD, and SNK to be excellent for protecting against high experimentwise type I error rates. In almost all of the simulations when at least some of the true means were not different, these three MC procedures had type I error rates around 5% or less, which corresponded well to the significance level at which the tests were conducted ($P = 0.05$). The other procedures had much higher experimentwise type I error rates, often greater than 40%.

Carmer and Swanson (2,3) found that type III errors (ie, reverse decisions) are rare with all of these MC procedures. Assuming no type III errors and nonzero true differences of the means, the correct-decision rate equals: 100 − (type II error rate), when all values are expressed on a percentage basis. The sensitivity of an MC procedure depends on its ability to correctly detect real differences among means (3,5,16). This sensitivity is represented by high correct-decision rates. The correct-decision rate depends on number of replications, number of treatments, magnitude of true relative differences, and level of homogeneity among the true means (3). Some observed correct-decision rates for Carmer and Swanson's (3) simulations of 10 treatment means with four replications are presented in Table 2. For small relative differences of the true means (d), all of the procedures have low correct-decision rates. As exemplified by these data, correct decisions increase with increases in magnitude of the true differences. The LSD, FLSD, DMRT, DLSD, and BLSD procedures consistently detected real differences more efficiently than SSD, TSD, and SNK. In general, none of these procedures resulted in a 100% correct-decision rate. Even at a reasonable magnitude of true relative differences of the means ($d = 2.0$), a factor out of the control of the investigator, the best procedures only made approximately 80% correct decisions in experimental designs with four replications. With six replications, the best procedures had correct-decision rates of approximately 90% (3).

The choice of MC procedure should depend on the costs attributable to type I and type II errors. If the commission of an experimentwise type I error is more serious (costly) than a type II

TABLE 1. Mean separation of 10 corn (*Zea mays* L.) cultivars by seven multiple comparison (MC) procedures and the Scott-Knott cluster analysis method

Cultivar	MDMV incidence[y]	MC Procedure[x]							
		FLSD	TSD	SSD	SNK	DMRT	DLSD	BLSD	Scott-Knott
1	1.000	a	a	a	a	a	a	a	a
2	0.993	a	a	a	a	a	ab	ab	a
3	0.969	ab	a	ab	a	ab	abc	abc	a
4	0.964	ab	a	ab	a	ab	abc	abc	a
5	0.929	ab	ab	abc	a	ab	bc	bc	a
6	0.905	b	ab	abc	a	b	c	c	a
7	0.813	c	bc	bc	b	c	d	d	b
8	0.785	c	c	c	b	c	d	d	b
9	0.775	c	c	c	b	c	d	d	b
10	0.108	d	d	d	c	d	e	e	c
Critical value		0.073	0.119	0.156	0.073–[z] 0.119	0.073–[z] 0.086	0.064	0.065	
No. of significant differences		29	25	19	27	29	30	30	

[x] FLSD = Fisher's least significant difference; TSD = Tukey's significant difference; SSD = Scheffe's significant difference; SNK = Student-Newman-Keuls; DMRT = Duncan's multiple range test; DLSD = Duncan's least significant difference; BLSD = Bayesian least significant difference. Means followed by the same letter are not significantly different at $P = 0.05$, or $k = 100$ for DLSD and BLSD (5,16).

[y] Mean proportion of plants infected by maize dwarf mosaic virus (MDMV) 63 days after planting in a field experiment with three replications in 1981.

[z] Critical value varies with range of means.

error, the SSD, TSD, or SNK procedure should be used. This situation may occur in a confirmatory experiment for which the investigator wishes to avoid having a "weakly" supported claim of treatment differences. If an investigator wishes to detect real differences among means (ie, commission of a type II error is more costly than a type I), then the LSD, FLSD, DMRT, DLSD, or BLSD could be used. By using the reasoning of Carmer and Swanson (3), a few of these procedures can be eliminated. The LSD should be eliminated because it performs no better than the FLSD and provides much poorer protection against type I errors. The DMRT could be eliminated because it consistently had a higher type II error rate than did the other four procedures. Since the DLSD is a less exact test than the BLSD, especially for small number of treatments and few degrees of freedom for the standard error (16), it could also be eliminated. This selection leaves the FLSD and BLSD procedures for multiple comparisons. In simulation studies, these two procedures produce very similar results (2,3). Those who prefer a simple test should choose the FLSD; those who prefer complex Bayesian arguments and loss functions should use the BLSD. Once an MC procedure is chosen, the investigator may wish to alter the significance level (P) of the test from the "standard" value of $P = 0.05$ in order to more thoroughly control the type I and type II error rates (1).

Cluster analysis. The MC procedures typically produce overlapping groups of means. In Table 1, for example, some means are followed by as many as three letters. With large numbers of treatments, means followed by 10 or more letters are not uncommon. Interpretation of these results often is difficult. It is sometimes advantageous to split treatments, especially cultivars, into nonoverlapping homogeneous groups. Scott and Knott (14) proposed a cluster analysis method for assigning means to distinct groups as a follow-up to ANOVA. Gates and Bilbro (6) illustrated the use of the Scott-Knott procedure for agronomic studies; their article should be consulted for details of the mean separation. This procedure first attempts to separate the means into two groups. The two groups are then tested separately for additional separations, and the partitioning is continued until groups of single means or groups of homogeneous means (or both) are found (6).

The groupings of the example means by using the Scott-Knott procedure are in the last column of Table 1. The last two groups agree with several of the MC procedures. The overlapping means of cultivars 1 through 6 with MC procedures resulted in one homogeneous group with the Scott-Knott method. Although in this example the Scott-Knott method produced the same separation of means as SNK, with other data the Scott-Knott produced results more similar to BLSD or DMRT (*unpublished*). We have found the Scott-Knott procedure to be a very useful alternative to the classical MC methods for grouping corn cultivars for susceptibility to MDMV.

The Scott-Knott method has not been subjected to the same degree of numerical analysis as the MC procedures, and thus the various error rates are not as well understood. In one simulation study, however, the experimentwise type I error rates for the Scott-Knott method were more than double the corresponding rates for FLSD (17). On the other hand, Scott-Knott and FLSD had approximately equal correct-decision rates with large relative differences of true means ($d > 2.0$). With small relative differences of the true means ($d < 2.0$), Scott-Knott had considerably higher correct-decision rates than FLSD (17).

Conclusions. Plant pathologists should be aware that in many experimental design situations other procedures are superior to MC for analyzing their data (4,7–11,13,15). When MC procedures are appropriate, care must be taken in choosing which of the MC methods to use. If the costs are greater if means are determined to be significantly different from each other when they are, in fact, equal, then SSD, TSD, or SNK should be used. In most situations, however, experiments are designed to detect treatment differences. To detect these differences, tests with high correct-decision rates should be used. Based on simulation studies (2,3), the FLSD or BLSD would seem to be appropriate choices. Researchers should discuss the costliness of the possible errors with a consulting statistician prior to designing an experiment or doing an analysis.

If nonoverlapping groups of means are desirable, cluster analysis techniques should be used. One of these procedures, the Scott-Knott, is a very useful follow-up to ANOVA. Cluster analysis should be especially useful for grouping large numbers of qualitative treatments and it could also be used in conjunction with one of the traditional MC procedures.

All of these methods are statistically valid for certain experiments, but that does not mean they have the same properties or produce the same results. Even the best procedures may result in many errors. Care must be taken when drawing conclusions from the results of any of these procedures.

TABLE 2. Correct-decision rates (%) of eight multiple comparison (MC) procedures for simulations of 10 treatment means with four replications[a]

MC procedure[c]	Relative difference of true means[b]	
	1.0	2.0
LSD	27.8	77.6
FLSD	27.7	77.3
TSD	3.1	28.9
SSD	0.4	7.9
SNK	9.4	48.5
DMRT	22.3	71.8
DLSD	34.1	81.7
BLSD	30.4	78.6

[a] Data from Carmer and Swanson (3).

[b] Relative difference $= (\mu_j - \mu_k)/\sigma$, in which μ_j and μ_k are the respective true means of treatments j and k, and σ is the standard deviation of the population.

[c] LSD = least significant difference; FLSD = Fisher's least significant difference; TSD = Tukey's significant difference; SSD = Scheffe's significant difference; SNK = Student-Newman-Keuls; DMRT = Duncan's multiple range test; DLSD = Duncan's least significant difference; and BLSD = Bayesian least significant difference.

LITERATURE CITED

1. Carmer, S. G. 1976. Optimal significance levels for application of the least significant difference in crop performance trials. Crop Sci. 16:95-99.
2. Carmer, S. G., and Swanson, M. R. 1971. Detection of differences between means: A Monte Carlo study of five pairwise multiple comparison procedures. Agron. J. 63:940-945.
3. Carmer, S. G., and Swanson, M. R. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J. Am. Stat. Assoc. 68:66-74.
4. Chew, V. 1976. Comparing treatment means: A compendium. HortScience 11:348-356.
5. Duncan, D. B. 1965. A Bayesian approach to multiple comparisons. Technometrics 7:171-222.
6. Gates, C. E., and Bilbro, J. D. 1978. Illustration of a cluster analysis method for mean separation. Agron. J. 70:462-465.
7. Gill, J. L. 1973. Current status of multiple comparison of means in designed experiments. J. Dairy Sci. 56:973-977.
8. Hicks, C. R. 1973. Fundamental Concepts in the Design of Experiments. Holt, Rinehart, and Winston, New York. 349 pp.
9. Johnson, S. B., and Berger, R. D. 1982. On the status of statistics in PHYTOPATHOLOGY. Phytopathology 72:1014-1015.
10. Little, T. M. 1981. Interpretation and presentation of results. HortScience 16:637-640.
11. Neter, J., and Wasserman, W. 1974. Applied Linear Statistical Models. Richard D. Irwin, Inc., Homewood, IL. 842 pp.
12. O'Neill, R., and Wetherill, G. B. 1971. The present state of multiple comparison methods. J. R. Stat. Soc. B 33:218-250.
13. Petersen, R. G. 1977. Use and misuse of multiple comparison procedures. Agron. J. 69:205-208.
14. Scott, A. J., and Knott, M. 1974. A cluster analysis method for grouping means in the analysis of variance. Biometrics 30:507-512.
15. Snedecor, G. W., and Cochran, W. G. 1967. Statistical Methods. Iowa State University Press, Ames. 593 pp.
16. Waller, R. A., and Duncan, D. B. 1969. A Bayes rule for the symmetric multiple comparisons problem. J. Am. Stat. Assoc. 64:1484-1503.
17. Willavize, S. A., Carmer, S. G., and Walker, W. M. 1980. Evaluation of cluster analysis for comparing treatment means. Agron. J. 72:317-320.