WILLIAM H. SWALLOW, Department of Statistics, North Carolina State University, Raleigh

Those Overworked and Oft-Misused Mean Separation Procedures—Duncan's, LSD, etc.

Mean separation or multiple comparison procedures are widely used in analyzing scientific data, usually as follow-up procedures after an analysis of variance has been performed. Once a significant F has indicated that a group of treatment means are not all equal, one naturally wishes to explore the treatment differences further. One way this is often done is with a mean separation procedure, usually by making pairwise comparisons of the treatment means in question.

The mean separation procedures most often used are Duncan's and Newman-Keuls' multiple range tests, the LSD (least significant difference), the HSD (Tukey's w or honestly significant difference), and Waller-Duncan's procedure (5). These procedures are used *far* more often than they ought to be, however. They are *not* all-purpose procedures for comparing means indiscriminately, nor were they ever intended to be. When Petersen (4) scanned the 1975 volume of the *Agronomy Journal*, he noted that 40% of the papers used a mean separation procedure (usually Duncan's). He concluded that 40% of those applications were "entirely inappropriate," 30% could have used a more suitable analysis, and only 30% used a mean separation procedure appropriately. Despite a number of papers on this subject (1–4), abuses of these procedures are still very easy to find.

So when *is* it inappropriate to use a mean separation procedure? The answer lies in considering the treatment design, by which I mean the nature of the treatments in the experiment and their interrelationships. Mean separation procedures were developed for cases where the treatment set lacked structure, that is, where the treatments were just a collection of varieties or perhaps chemicals with no particular interrelationships. Most treatment designs are not of this type. Usually, the treatment set has a structure, and the statistical analysis should recognize that structure. When that structure is ignored in the statistical analysis, as it is when a mean separation procedure is used to make all pairwise comparisons, then the statistical analysis will not be the best (most pertinent) analysis and may be entirely inappropriate.

The following examples provide a basis for discussion of the most common misapplications of mean separation procedures. For verisimilitude, examples 1 and 2 are closely based on misapplications published recently, but the data have been altered to obviate citing specific papers for abuses that are widespread.

Example 1. Quantitative treatments. Perhaps the most glaring abuse of a mean separation procedure is using it on a gradient treatment design, that is, a set of treatments that are increasing "dosages" of a quantitative factor. Examples of such treatments include dosages or concentrations of a chemical treatment, row spacings, times of application, and temperatures. That the levels or dosages may be planned, not random, is seldom relevant.

Table 1 illustrates a possible presentation associated with this misuse of a mean separation procedure. To ask whether the first treatment level differs from the second, then from the third, then from the fourth, etc., by making all pairwise comparisons of means, as is done in Table 1, ignores the logic of the treatment design. The focus of a gradient treatment design is to investigate the "dose-response" relationship. To do that, one should plot the response (Y) against the treatment level (X) and look for an equation describing the relationship between Y and X. If theory suggests a meaningful mathematical form for that equation, then fitting an equation of that form is preferable. Otherwise (usually), one merely tries to find a simple equation that fits the data reasonably well. Polynomials are popular for their ease of use and ability to fit a wide variety of data. For this example, the quadratic equation

$$\text{Yield} = 4{,}025.3 + 1{,}478.3(\text{Rustkill}) - 349.8(\text{Rustkill})^2$$

accounts for over 98% of the treatment sum of squares. (A quadratic equation fit the real data on which this example was based even better!) This equation not only provides a compact summary of the dose-response relationship (over the range of Rustkill rates in the data—beware of extrapolation!), but also allows prediction of wheat yield at treatment levels not included in the data. For example, for Rustkill applied at 1.15 kg/ha, the predicted wheat yield is 5,263 kg/ha. Having an equation for the dose-response relationship also can be helpful in estimating the point (threshold) at which treatment becomes cost-effective or the treatment level associated with a maximum or minimum response.

So, for quantitative treatments, estimating the dose-response relationship (or, in higher dimensions, the response surface) through curve fitting is appropriate. Pairwise comparison of the treatment means is not likely to shed much light on the dose-response relationship. As Little (3) aptly noted, "Perhaps it is fortunate that Galileo did not have Duncan's test at his disposal, for he might have failed to come up with the beautifully simple equation, v = gt."

Example 2. Factorial experiments. Factorial treatment designs are common and are widely recommended for experiments designed to investigate possible interactions of factors. The treatment set for a two-factor factorial can be displayed in a two-way table (rows and columns), highlighting the key point that the treatments derive from a "crossing" of the levels of factor A with those of factor B; a k-factor factorial can be displayed similarly with a k-way table.

The cross-classificational nature of a factorial treatment design should not be ignored in the statistical analysis. Thus, with a factorial it is almost always wrong to use a mean separation procedure on the full set of treatments. That notwithstanding, one often sees the sort of analysis presented in Table 2. Only the most astute reader will gain any understanding of the main effects of the

Table 1. Example 1: Effect on wheat yield of leaf rust treatment with different rates of Rustkill[y]—a flawed analysis and presentation

Treatment and rate/ha	Yield (kg/ha)
Control (0 kg)	4,134 e[z]
Rustkill 25W 0.25 kg	4,232 e
Rustkill 25W 0.50 kg	4,635 d
Rustkill 25W 0.75 kg	4,965 c
Rustkill 25W 1.00 kg	5,199 b
Rustkill 25W 1.25 kg	5,311 b
Rustkill 25W 1.50 kg	5,505 a
Rustkill 25W 2.00 kg	5,551 a
LSD ($P = 0.05$)	125

[y] Not real data.
[z] Means followed by the same letter are not significantly different.

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.

nematicide and herbicide treatments and of their interaction from this analysis with Duncan's test. Indeed, most readers will fail even to recognize the factorial nature of the treatment set.

For the same eight treatments, Table 3 makes the factorial treatment design explicit and shows the appropriate partitioning of the treatment sum of squares (ie, that suggested by the treatment design) into pieces reflecting the effects of presence vs. absence of the nematicide (rows), differential effects of the herbicides including none (columns), and the interaction of nematicide and herbicide treatments. Over 99% of the treatment sum of squares is attributable to herbicide differences; the main effect for nematicide and interaction are not significant.

Although it was inappropriate to apply any mean separation procedure to the full factorial set of eight treatments, it does seem appropriate to compare the four herbicide treatments using a mean separation procedure as done on the column means in the two-way table of Table 3. It seems appropriate because I think the experimenter would want to make all possible pairwise comparisons of these four treatments (cf example 3). The main effect (column) means are used because there was no significant interaction. If the interaction had been significant, I would have compared the four herbicide means within each level of the other factor (ie, within each row of the two-way table). In contrast to the muddled message in Table 2, inferences flow straightforwardly from Table 3: Peach tree growth was unchanged with use of Nemakill; all three herbicides increased yield significantly but the increase with Goal was significantly greater than with either Surflan or Solicam; there was no significant interaction of the herbicide and nematicide treatments. The power gained in comparing herbicide treatments averaged across nematicide treatments, exploiting the factorial's "hidden replication," separated Goal from Solicam, a difference not evident in Table 2.

Table 2. Example 2: Effect on new growth of peach trees of nematicide and herbicide treatments for *Pratylenchus penetrans* and weeds[y]—a flawed analysis and presentation

Treatment and rates/acre	New growth (cm)
Control	55.9 cd[z]
Nemakill 15G (133 lb)	50.8 d
Goal 2E (1 gal)	180.8 a
Surflan 4AS (1 gal)	109.6 bc
Solicam 80W (5 lb)	137.1 ab
Nemakill 15G (133 lb) + Goal 2E (1 gal)	190.3 a
Nemakill 15G (133 lb) + Surflan 4AS (1 gal)	94.8 bcd
Nemakill 15G (133 lb) + Solicam 80W (5 lb)	137.9 ab

[y] Not real data.
[z] Means followed by the same letter are not significantly different ($P = 0.05$) according to Duncan's multiple range test.

Example 3. Contrasts and preplanned tests. Many treatment sets incorporate a structure that strongly suggests the treatments were selected with particular comparisons in mind. Often the treatments fall into natural subgroups that "cry out" for comparison. Table 4 shows one such treatment set from Steel and Torrie (5, pp. 205-208) and the comparisons or contrasts that follow naturally from the treatment design. Using the method of orthogonal contrasts, the sum of squares for treatments with seven degrees of freedom can be partitioned into single-degree-of-freedom sums of squares to test the seven pertinent questions listed in Table 4; Steel and Torrie (5) provide the details. Note that some of these contrasts are not pairwise; for example, the second compares a group of two treatments vs. a group of five. Some of the mean separation procedures can also do nonpairwise comparisons, but they are rarely used that way.

When relevant hypotheses follow from the treatment design, as do the seven in this example and as did the tests for main effects and interaction in example 2, the overall F test is not prerequisite, relevant, or recommended. In fact, a nonsignificant overall F may wrongly dissuade the experimenter from testing the preplanned hypotheses of interest; when most of the treatments differ little, the overall F may fail to detect that some differences do exist.

It should be said that relevance is far more important than orthogonality. When the treatment design suggests nonorthogonal contrasts, so be it. The mathematical niceties of orthogonality are far less important than extracting all pertinent information from the data.

Whereas the misuses of mean separation procedures illustrated in examples 1 and 2 seem to me incontrovertible, there is more room for judgment in deciding what is preplanned and should therefore be tested with contrasts rather than a mean separation procedure. I applied the LSD to the four herbicide treatment means in Table 3, feeling that the structure in that group of four treatments was minimal. Someone else might have argued that Goal and Solicam were more similar to each other (eg, in chemical structure and mode of application) than to Surflan, so one should instead have calculated three contrasts: control vs. herbicide, Surflan vs. Goal and Solicam, and Goal vs. Solicam. At the extreme, there are statisticians who argue that everything should be viewed as preplanned; that if it doesn't seem so, it's because the treatment set was poorly designed. Those statisticians would cheerfully dispense with mean separation procedures altogether.

It could be said that real life is more complicated than examples 1, 2, and 3—that treatment sets are usually more complex. That may well be true, but two points come to mind. First, a more complex set of treatments may mean that the analysis will be more complex but doesn't void any of the arguments made

Table 3. Example 2: Factorial structure and partitioning

	Herbicide				
Nematicide	**None**	**Goal**	**Surflan**	**Solicam**	**Mean**
None	55.9	180.8	109.6	137.1	120.9
Nemakill	50.8	190.3	94.8	137.9	118.5
Mean	53.4 a[z]	185.6 c	102.2 b	137.5 b	

Source of variation	df	Sum of squares
Treatments	7	18,891.2
Nematicide	1	11.5
Herbicide	3	18,723.3
Interaction	3	156.5

[z] Herbicide means followed by a common letter are not significantly different (LSD = 39.4, $P = 0.05$).

Table 4. Example 3: Treatments for corn seedlings infected with *Diplodia* spp. and implied contrasts of interest

Treatments		
A	=	untreated control
B,C	=	mercuric fungicides
D,H	=	nonmercuric fungicides, company I
E,F,G	=	nonmercuric fungicides, company II (F,G are newer formulations of E)

Implied contrasts	
1. Control vs. treated	(A vs. rest)
2. Mercuric vs. nonmercuric	(B,C vs. D,E,F,G,H)
3. Comparing mercurics	(B vs. C)
4. Company I vs. company II	(D,H vs. E,F,G)
5. Comparing products, company I	(D vs. H)
6. Old vs. new formulations, company II	(E vs. F,G)
7. Comparing new formulations, company II	(F vs. G)

here. If 14 treatments include a 3 × 4 factorial set plus two miscellaneous treatments, the factorial part should be analyzed as a factorial. The presence of odd treatments doesn't convey license to ignore the rest of the structure in the treatment set and proceed with Duncan's test. And second, a hodgepodge treatment set often suggests that the experimental objectives were not well thought out.

In judging whether a mean separation procedure has been used improperly, experimental design is irrelevant. It is immaterial whether the experiment was run as a completely randomized design, a randomized complete block design, or a split plot design. What counts is the nature of the treatments, that is, the treatment design.

I think mean separation procedures do have a place in data analysis, despite their frequent misuse. So, assuming it is appropriate to use one, which procedure should one choose? There is room for differing opinions. Very briefly, here are some of my own feelings. First, I would never use a multiple range test (Duncan's or Newman-Keuls'). In using a multiple range test, means are ranked and then compared by one statistic if they are adjacent in the ranked list, by another statistic if they are separated by one mean, by yet another if they are separated by two means, etc. Why should my perception of a difference between treatments A and B depend on whether the other treatments in the experiment happened to give means that fell between those for A and B? Furthermore, since these procedures differ fundamentally in the meaning they attach to the error rate, I prefer procedures that define the error rate in easy-to-describe ways (LSD and Tukey's HSD). And, most importantly, multiple range tests do not lend themselves to easy construction of sets of simultaneous confidence intervals. Interval estimation is far more informative than hypothesis testing, ought to be used more often, and is easily done with LSD, HSD, or the Waller-Duncan significant difference.

Second, unless one has very few treatments, the HSD and Scheffé's test are too conservative for most applications. They offer so much protection against type I errors (false positives: claiming differences that are not real) that it is difficult to find any treatment differences, and type II errors (false negatives: failing to detect real differences) become too likely.

Third, I usually choose the LSD or the Waller-Duncan test. It is well known that the LSD is prone to type I errors, but if one requires a significant F (evidence that treatment differences do exist) before applying the LSD, then the risk of type I errors seems acceptable; this is often called using the "protected" LSD. The Waller-Duncan test is conceptually appealing; the value of the statistic falls somewhere between the LSD and HSD according to the calculated F. When the F is small (little evidence of treatment differences), the Waller-Duncan statistic is close to the HSD, providing a high level of protection against type I errors. When the F is large, it approaches the LSD, making it easier to identify treatment differences that the F has indicated do exist. However, the meaning of the error rate for the Waller-Duncan test is not easily described, the statistic is more complicated, and the test suffers from limited availability of tables.

Which mean separation procedure one elects to use—when it is appropriate to use one—is *far* less important than knowing when they are *all inappropriate*. The key to deciding when they are all inappropriate lies in the treatment design.

LITERATURE CITED

1. Chew, V. 1976. Comparing treatment means: A compendium. HortScience 11:348-357.
2. Chew, V. 1980. Testing differences among means: Correct interpretation and some alternatives. HortScience 15:467-470.
3. Little, T. M. 1978. If Galileo published in *HortScience*. HortScience 13:504-506.
4. Petersen, R. G. 1977. Use and misuse of multiple comparison procedures. Agron. J. 69:205-208.
5. Steel, R. G. D., and Torrie, J. H. 1960. Principles and Procedures of Statistics. 2nd ed. McGraw-Hill, New York.

Dr. Swallow, associate professor in the Department of Statistics at North Carolina State University, Raleigh, contributed this article at the invitation of D. F. Ritchie, Editor, Fungicide and Nematicide Tests, *published annually by the New Fungicide and Nematicide Data Committee of The American Phytopathological Society. Copies of current and past volumes may be obtained from Richard E. Stuckey, Business Manager F & N Tests, Plant Pathology Department, University of Kentucky, Lexington 40546.*