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.

© 1984 The American Phytopathological Society

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.

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 2. Example 2: Effect on new growth of peach trees of nematicide and herbicide treatments for *Pratylenchus penetrans* and weeds[1]—a flawed analysis and presentation

Treatment and rates/acre	New growth (cm)
Control	55.9 cd[2]
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

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

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[1]	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

[1] 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.*

Order your copy today!

Phytophthora:

Its Biology, Taxonomy, Ecology, and Pathology

Edited by D. C. Erwin, S. Bartnicki-Garcia, and P. H. Tsao, University of California at Riverside.

This book is the first comprehensive account of *Phytopthora*—a fungus of worldwide distribution affecting agronomic and horticultural crops, and woody plants. It includes current research as well as in-depth reviews of almost all the major areas of knowledge related to *Phytophthora*. It will be a valuable reference for years to come.

Prices include postage and handling for the United States. We accept MasterCard, Visa, and American Express.

$59.00 APS members $68.00 Nonmembers

Order by phone. It's easy. It's fast. It's free.

Now you can order APS books over the telephone from anywhere in the continental United States except Minnesota. Simply dial our no-charge number between the hours of 8:00 a.m. and 4:00 p.m. (Central Standard Time), Monday through Friday (excluding holidays).

1-800-328-7560

Minnesota residents call

(612) 454-7250

gaged one of the other two. This resulted in an immediate change in the size of sprocket driving the planter boxes and thereby changed the seeding rate.

The described electric clutch assembly was connected by a roller-chain drive to the original manually adjusted seeding control unit. The combination provides an almost unlimited choice in seeding rates. The device shown in Fig. 1 consists of 34-, 24-, and 12-tooth sprockets on the shaft assemblies driven by the clutches. Therefore, the switching of the clutches changes the seeding rates in roughly ã 3 to 2 to 1 ratio. This reflects the desired cut in seeding rate we wanted to achieve for our water management study. To obtain other ratios, different size sprockets could be selected. For example, if it was desirable to reduce the seeding rate by only about 10 or 20%, sprockets could be selected to make this change. With this equipment, the range in seeding rate can be from 8,000 to 400,000 seed/ha. The device described allows choosing one of three seeding rates from the tractor seat, but it could be built for two, and possibly for even more than three. However, it may not be practical to have more than three clutches running in sequence.

The device is ideally suited for changing seeding rates. The same concept, however, could be used to change fertilizer or pesticide rates in instances where the materials are applied by chain- or belt-driven assemblies. The concept could also be used to allow the flexibility of fertilizing only portions of a field, in the event there were isolated areas where a specific plant nutrient was required.

The device can be readily constructed at moderate cost. The cost of materials used for the device described here was approximately $600.00.

Acknowledgments

The author is indebted to Charles Wright, farmer, Panhandle, Tex., for the original idea; to Frank Wood and Odis Henry, farm manager and heavy mobile equipment mechanic, respectively, USDA-ARS, Bushland, Tex., for constructing and testing the device; and to Thomas Marek, agricultural engineer, The Tex. Agric. Exp. Stn., Bushland, Tex., for helpful suggestions.

Literature Cited

Stewart, B. A., J. T. Musick, and D. A. Dusek. 1981. A management system for the conjunctive use of rainfall and limited irrigation of graded furrows. Soil Sci. Soc. Am. J. 81:413–419.

ANALYSIS OF COMBINED EXPERIMENTS[1]

M. S. McIntosh[2]

Abstract

Most field experiments are conducted over two or more locations or years, yet statistical references do not contain sufficient detail for complete analysis. The purpose of this paper is to provide a reference for analysis of combined experiments. Tables include sources of variation, degrees of freedom, and F-ratios for one factor and split-plot experiments combined over locations and/or years. The F-ratios are given for fixed, mixed, and random models.

Additional index words: **Experimental design, Series of experiments, Statistics.**

Most field experiments are conducted over two or more locations or years, yet there is no standard reference which provides all of the details necessary for combining analyses of experiments with more than one factor. Two of the references (Steel and Torrie, 1980; Little and Hill, 1978) most commonly used by agronomists do not contain analyses for combining analyses of annual crops. Kempthorne (1952), Cochran and Cox (1957), and Snedecor and Cochran (1967) describe the procedure for combined analysis of one factor experiments, but do not describe the test of the average response to treatments over years or locations. The test of the main effect of locations or years may be of interest to researchers, but is not readily available in the literature. Although, the appropriate tests of years and locations are relatively straight forward to derive, many agronomists might not have the statistical skills to identify all sources of variation and derive their expected mean squares. It is important to completely define the statistical model even if a researcher is not interested in testing the main effects of years or locations. To correctly analyze an experiment all terms must be accounted for because most computer software packages automatically pool any unaccounted sources of variation with the error term. The result of an under-defined model could be an inflated error term.

[1] Scientific Article No. A-3167 and Contribution No. 6236 of the Univ. of Maryland Agric. Exp. Stn. Contribution from Dep. of Agronomy, Univ. of Maryland, College Park 20742. Received 17 May 1982.

[2] Assistant professor, Dep. of Agronomy, Univ. of Maryland, College Park, MD 20742.

Table 2. F-ratios used to test effects for randomized complete block experiments combined over locations.

Sources of variation	Mean squares	F-tests		
		RL-RT†	RL-FT	FL-FT
Locations	M_1	$(M_1+M_4)/(M_2+M_4)$	M_1/M_2	M_1/M_2
Blocks/Locations	M_2			
Treatment	M_3	M_3/M_4	M_3/M_4	M_3/M_4
Location × Treatment	M_4	M_4/M_4	M_4/M_4	M_4/M_4
Pooled error	M_4			

† R = random, F = fixed, L = location, T = treatment.

Table 1. Expected mean squares for randomized complete blocks experiments combined over locations.

Sources of variation	df	Mean squares	Expected mean squares		
			RL-RT†	RL-FT	FL-FT
Locations	$l-1$	M_1	$\sigma_\epsilon^2 + r\sigma_{TL}^2 + t\sigma_{R(L)}^2 + rt\sigma_L^2$	$\sigma_\epsilon^2 + t\sigma_{R(L)}^2 + rt\sigma_L^2$	$\sigma_\epsilon^2 + t\sigma_{R(L)}^2 + rt\theta_L^2$
Blocks/locations	$l(r-1)$	M_2	$\sigma_\epsilon^2 + t\sigma_{R(L)}^2$	$\sigma_\epsilon^2 + t\sigma_{R(L)}^2$	$\sigma_\epsilon^2 + t\sigma_{R(L)}^2$
Treatment	$t-1$	M_3	$\sigma_\epsilon^2 + r\sigma_{TL}^2 + rl\sigma_T^2$	$\sigma_\epsilon^2 + r\sigma_{TL}^2 + rl\theta_T^2$	$\sigma_\epsilon^2 + rl\theta_T^2$
Location × treatment	$(l-1)(t-1)$	M_4	$\sigma_\epsilon^2 + r\sigma_{TL}^2$	$\sigma_\epsilon^2 + r\sigma_{TL}^2$	$\sigma_\epsilon^2 + r\theta_{TL}^2$
Pooled error	$l(r-1)(t-1)$	M_4	σ_ϵ^2	σ_ϵ^2	σ_ϵ^2

† R = random, F = fixed, L = location, T = treatment.

Table 3. F ratios used to test effects in a randomized complete block experiment combined over years and locations.

Sources of variation	df	Mean squares	F-tests RY-RL-RT†	RY-RL-FT	RY-FL-RT	FY-RL-FT	RY-FR-FT
Years	$y-1$	M_1	$(M_1+M_6)/(M_4+M_6)$	M_1/M_5	$(M_1+M_6)/(M_4+M_6)$	M_1/M_5	M_1/M_4
Locations	$l-1$	M_2	$(M_2+M_6)/(M_3+M_7)$	M_2/M_5	$(M_2+M_6)/(M_3+M_7)$	M_2/M_4	M_2/M_4
Years × Locations	$(y-1)(l-1)$	M_3	$(M_3+M_6)/(M_4+M_6)$	M_3/M_4	$(M_3+M_6)/(M_4+M_6)$	M_3/M_4	M_3/M_4
Blocks/Years × Locations	$(r-1)yl$	M_4					
Treatments	$t-1$	M_5	$(M_5+M_6)/(M_6+M_7)$	$(M_5+M_6)/(M_6+M_7)$	M_5/M_6	M_5/M_7	M_5/M_6
Treatments × Years	$(t-1)(y-1)$	M_6	M_6/M_6	M_6/M_6	M_6/M_6	M_6/M_6	M_6/M_6
Treatments × Locations	$(t-1)(l-1)$	M_7	M_7/M_6	M_7/M_6	M_7/M_6	M_7/M_6	M_7/M_6
Treatments × Years × Locations	$(t-1)(y-1)(l-1)$	M_8	M_8/M_6	M_8/M_6	M_8/M_6	M_8/M_6	M_8/M_6
Pooled error	$(r-1)(t-1)yl$	M_9					

† R = random, F = fixed, Y = years, L = locations, T = treatments.

Expected mean squares must be known in order to determine the F-ratios used to test the hypotheses of interest. Rules for determining expected mean squares for balanced data are described by Schultz (1955) and in many statistical texts (Kempthorne, 1952; Steel and Torrie, 1980; LeClerg et al., 1962; Snedecor and Cochran 1967; and Bennett and Franklin, 1954). The expectations for the unbalanced case do not always equal those of the balanced case (Hartley and Searle, 1969), depending on the variance and covariance definitions in the model. Hocking (1973) described three models with different variance and covariance definitions for the unbalanced case. The expectations used in the present paper to determine F-tests correspond to Hocking's Model I and are for the balanced case.

Computer programs will also generate expected mean squares. The Statistical Analysis System (SAS) includes the procedures VARCOMP and GLM with the RANDOM option which give the coefficients for and estimate variance components of mean squares. These procedures will produce the same expected mean squares as Hocking's Model III (Freund and Littell, 1981). For mixed models, the expectations generated by SAS differ from the balanced case expectations.

The purpose of this paper is to present some complete analysis of variance tables for combining balanced experiments that can be used by researchers as a reference to quickly and correctly identify sources of variation and the appropriate F-ratios.

Table 4. F-ratios used to test effects for split plot experiments arranged in a randomized complete block design and combined over locations, A and B fixed.

Sources of variation	df	Mean squares	F ratios FL†	RL
Locations	$l-1$	M_1	M_1/M_6	M_1/M_2
Blocks/Locations	$(r-1)l$	M_2		
A	$a-1$	M_3	M_3/M_6	M_3/M_4
A × Location	$(a-1)(l-1)$	M_4	M_4/M_6	M_4/M_6
Pooled error a	$(a-1)(r-1)l$	M_5		
B	$b-1$	M_6	M_6/M_{10}	M_6/M_7
B × Location	$(b-1)(l-1)$	M_7	M_7/M_{10}	M_7/M_{10}
A × B	$(a-1)(l-1)$	M_8	M_8/M_{10}	M_8/M_9
A × B × Location	$(a-1)(b-1)(l-1)$	M_9	M_9/M_{10}	M_9/M_{10}
Pooled error b	$al(r-1)(l-1)$	M_{10}		

† F = fixed, R = random, L = location.

Results and Discussion

The expected mean squares and sources of variation for randomized complete block experiments combined over locations are given in Table 1. Table 1 can also be used for experiments combined over years by replacing years with locations. The appropriate F-tests differ depending on whether the locations and treatments are fixed or random effects (Table 2). In the case where both locations and treatments are random, there is no exact test for location effects. Therefore, an approximation is proposed

Table 5. F-ratios used to test effects of split plot experiments arranged in a randomized complete block design combined over location, A and B fixed.

Sources of variation	df	Mean squares	F-tests RY-RL†	RY-FL	FY-FL
Years	$y-1$	M_1	M_1/M_9	M_1/M_4	M_1/M_4
Locations	$l-1$	M_2	M_2/M_9	M_2/M_5	M_2/M_4
Years × Locations	$(y-1)(l-1)$	M_3	M_3/M_4	M_3/M_4	M_3/M_4
Blocks/Locations × Years	$yl(r-1)$	M_4			
A	$a-1$	M_5	$(M_5+M_9)/(M_6+M_7)$	M_5/M_6	M_5/M_6
A × Years	$(a-1)(y-1)$	M_6	M_6/M_9	M_6/M_6	M_6/M_6
A × Locations	$(a-1)(l-1)$	M_7	M_7/M_9	M_7/M_6	M_7/M_6
A × Years × Locations	$(a-1)(y-1)(l-1)$	M_8	M_8/M_9	M_8/M_6	M_8/M_6
Pooled error a	$yl(a-1)(r-1)$	M_9			
B	$(b-1)$	M_{10}	$(M_{10}+M_{12})/(M_{11}+M_{13})$	M_{10}/M_{11}	M_{10}/M_{10}
B × Years	$(b-1)(y-1)$	M_{11}	M_{11}/M_{13}	M_{11}/M_{10}	M_{11}/M_{10}
B × Locations	$(b-1)(l-1)$	M_{12}	M_{12}/M_{13}	M_{12}/M_{10}	M_{12}/M_{10}
B × Years × Locations	$(b-1)(y-1)(l-1)$	M_{13}	M_{13}/M_{13}	M_{13}/M_{10}	M_{13}/M_{10}
A × B	$(a-1)(b-1)$	M_{14}	$(M_{14}+M_{17})/(M_{15}+M_{16})$	M_{14}/M_{10}	M_{14}/M_{10}
A × B × Years	$(a-1)(b-1)(y-1)$	M_{15}	M_{15}/M_{17}	M_{15}/M_{10}	M_{15}/M_{10}
A × B × Locations	$(a-1)(b-1)(l-1)$	M_{16}	M_{16}/M_{17}	M_{16}/M_{17}	M_{16}/M_{10}
A × B × Years × Locations	$(a-1)(b-1)(y-1)(l-1)$	M_{17}	M_{17}/M_{10}	M_{17}/M_{10}	M_{17}/M_{10}
Pooled error b	$yla(b-1)(r-1)$	M_{10}			

† R = random, F = fixed, Y = years, L = location.

(Satterthwaite, 1946) using N_1' and N_2' degrees of freedom (df), where:

$$N_1' = (M_1 + M_5)^2/[M_1^2/(l-1) + M_5^2/(l)(r-1)(t-1)]$$

$$N_2' = (M_2 + M_4)^2/[M_2^2/l(r-1) + M_4^2/(l-1)(t-1)]$$

Before calculating the approximate df, time, and effort could be saved by testing the significance of locations using the df for locations for the numerator and the df for blocks/locations or locations × treatment (whichever is smaller) for the denominator. This is a conservative test and if the F is significant, it is not necessary to calculate N_1' and N_2'. If the F is not significant, using N_1' and N_2' will lower the critical F- value, possibly causing the F to become significant.

Table 3 gives the approximate F-tests to be used for combining randomized complete block experiments over locations and years. The F-ratios have been given for fixed, mixed, and random models. The F-ratios given for random years and fixed locations, can also be used for fixed years and random locations if years and location are switched under sources of variation. As in the previous example, approximate F-tests are appropriate for some of the comparisons.

Table 4 gives the analysis of variance to be used for combining split plot experiments over locations and years. The table includes the F-tests to be used for fixed or random locations and fixed treatment effects. Table 5 is an extension of Table 4 which also includes years as fixed or random effects.

Literature Cited

1. Bennett, C. A., and N. L. Franklin. 1954. Statistical analysis in chemistry and the chemical industry. John Wiley and Sons, New York.
2. Cochran, W. G., and G. M. Cox. 1957. Experimental designs. John Wiley and Sons, New York.
3. Freund, R. J., and R. C. Littell. 1981. SAS for linear models. SAS Institute Inc. Cary, N.C.
4. Hartley, H. O., and S. R. Searle. 1969. A discontinuity in mixed model analysis. Biometrics 25:573–576.
5. Hocking, R. R. 1973. A discussion of the two-way mixed model. Am. Stat. 27:148-152.
6. Kempthorne, O. 1951. Design and analysis of experiments. John Wiley and Sons, New York.
7. LeClerg, E. L., W. H. Leonard, and A. G. Clark. 1962. Field plot technique. Burgess Publishing Co., Minneapolis.
8. Little, T. M., and F. J. Hills. 1978. Agricultural experimentation. John Wiley and Sons, New York.
9. Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biom. Bull. 2:110-114.
10. Schultz, E. F., Jr. 1955. Rules of thumb for determining expectations of mean squares in analysis of variance. Biometrics 11:123-125.
11. Snedecor, G. W., and W. G. Cochran, 1967. Statistical methods. Iowa State Univ. Press, Ames.
12. Steel, R. G., and J. H. Torrie. 1980. Principles and procedures of statistics. McGraw-Hill Book Co., New York.

NEW BOOKS RECEIVED

METHODS AND PRINCIPLES OF MYCORRHIZAL RESEARCH—Edited by *N. C. Schenck*. American Phytopathological Society, 3340 Pilot Knob Road, St. Paul, MN 55121. 1982. 244 p. $22.00, member; $24.00, nonmember.

LABORATORY MANUAL FOR INTRODUCTORY SOIL SCIENCE, Sixth Edition—*H. D. Foth, L. V. Withee, H. S. Jacobs,* and *S. J. Thien*. Wm. C. Brown Company Publishers, 2460 Kerper Boulevard, Dubuque, IA 52001. 1982. 143 p.

FLEXCROP: A DRYLAND SYSTEMS MODEL—*A. D. Halvorson* and *P. O. Kresge*. USDA Production Res. Rep. No. 180. USDA, ARS, Washington, DC 20250. 1982. 49 p.

DESCRIPTION AND CULTURE OF LENTILS—*R. J. Summerfield, F. J. Muehlbauer,* and *R. W. Short*. USDA, Production Res. Rep. No. 181. USDA, ARS, Washington, DC 20250. 1982. 22 p.

AGRICULTURAL JOURNAL TITLES AND ABBREVIATIONS. The Oryx Press, 2214 North Central at Encanto, Phoenix, AZ 85004. 1982. 154 p.

SOIL MOISTURE AND TEMPERATURE REGIMES IN AFRICA—*A. Van Wambeke*. USDA, Soil Conservation Service, P.O. Box 2890, Washington, DC 20013. 1982.

A GUIDE FOR ENHANCEMENT OF FISH AND WILDLIFE ON ABANDONED MINE LANDS IN THE EASTERN UNITED STATES—*J. R. Nawrot, A. Woolf,* and *W. D. Klimstra*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402. 1982. 101 p.

CONTROLLED ATMOSPHERES FOR STORAGE AND TRANSPORT OF PERISHABLE AGRICULTURAL COMMODITIES—Edited by *D. G. Richardson* and *M. Meheriuk*. Timber Press, P.O. Box 1631, Beaverton, OR 97075. 1982. 390 p. $39.95.

METABOLISM OF HERBICIDES IN HIGHER PLANTS—*K. K. Hatzios* and *D. Penner*. Burgess Publishing Company, 7108 Ohms Lane, Minneapolis, MN 55435. 1982. 142 p. $29.95.

THE 5-YEAR OUTLOOK ON SCIENCE AND TECHNOLOGY (1981), including Source Materials, Vol. 1 and 2—National Science Foundation. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

SOILS OF THE GREAT PLAINS—*A. R. Aandahl*. University of Nebraska Press, 901 North 17th Street, Lincoln, NE 68588. 1982. 281 p. $28.50.

ASIAN CROPPING SYSTEMS RESEARCH: MICROECONOMIC EVALUATION PROCEDURES—*Gordon R. Banta*. International Development Research Centre, Box 8500, Ottawa, Canada K1G 3H9. 1982. 56 p. $5.00.

DIE GRASER ALS KULTURPFLANZEN UND UNKRAUTER AUF WIESE, WEIDE UND ACKER—*Asmus Petersen*. Adademie-Verlag, DDR—1080 Berlin, Leipziger Strasse 3–4. 1981. 280 p. (in German).

PRINCIPLES OF PLANT PATHOLOGY—*J. G. Manners*. Cambridge University Press, 32 East 57th Street, New York, NY 10022. 1982. 264 p. Hardcover, $47.50; paperback, $17.95.

HORTICULTURAL REVIEWS, Vol. 4—Edited by *Jules Janick*. AVI Publishing Co., Inc., P.O. Box 831, Westport, CT 06881. 1982. 281 p. $33.00.

AGRICULTURAL DEVELOPMENT PLANNING IN THAILAND—Edited by *Kenneth J. Nicol, Somnuk Sriplung,* and *Earl O. Heady*, Iowa State University Press, South State Ave., Ames, IA 40010. 1982. 326 p. $21.50.

AGRICULTURAL RESEARCH POLICY—*Vernon W. Ruttan*. Univ. of Minnesota Press, 2037 University Avenue Southeast, Minneapolis, MN 55414. 1982. 369 p. Cloth, $32.50; paper, $13.95.

APPLICATION OF ELECTRO-ULTRAFILTRATION (EUF) IN AGRICULTURAL PRODUCTION—Edited by *K. Nemeth*. Martinus Nijhoff Publishers BV, P.O. Box 566, 2501 CN The Hague, The Netherlands. 1982. 138 p. US$22.00.

A LONG DEEP FURROW: Three Centuries of Farming in New England—*Howard S. Russell*; Abridged edition by Mark Lapping. The University Press of New England, Box 979, Hanover, NH 03755. 1982. 380 p. $12.95.

PROCEEDINGS OF THE NATIONAL SEMINAR ON GINGER AND TURMERIC, Calicut, India, 8–9 April 1980. Edited by *M. K. Nair, T. Premkumar, P. N. Ravindran* and *Y. R. Sarma*. Central Plantation Crops Research Institute, Regional Station, Post Bag No. 1, Marikunnu P.O., Calicut—673 012, Kerala, India. 1982. 250 p. Rs. 60/—.

ENVIRONMENT AND PLANT ECOLOGY, Second Edition—*John R. Etherington*. John Wiley & Sons, Inc., One Wiley Drive, Somerset, NJ 08873. 1982. 487 p. $59.95.

NITROGEN CYCLING IN SOUTH-EAST ASIAN WET MONSOONAL ECOSYSTEMS—Edited by *R. Wetselaar, J. R. Simpson* and *T. Rosswall*. Australian Academy of Science, P.O. Box 783, Canberra City ACT 2601 Australia. 1981. 216 p. $25.00 plus postage.

RESOURCE CONSERVATION GLOSSARY, Third Edition. Soil Conservation Society of America, 7515 Northeast Ankeny Road, Ankeny, IA 50021. 1982. 193 p. $7.00; SCSA members, $6.00.

FOOD, NUTRITION AND CLIMATE—Edited by *Sir Kenneth Blaxter* and *Leslie Fowden*. Applied Science Publishers Ltd., 22 Ripplesside Commercial Estate, Ripple Road Barking Essex 1G110SA, England. 1982. 422 p. $72.00.

Baby Bear's Dilemma: A Statistical Tale[1]

S. G. Carmer and W. M. Walker[2]

ABSTRACT

An allegorical and satirical, but also, we hope an accurate and humorous expository look at the problem researchers face in choosing a pairwise multiple comparisons procedure for detecting differences among treatment means. The primary objective is to present, from several points of view, some of the arguments and resulting confusion surrounding the use of the least significant difference vis-a-vis Tukey's *w* procedure or honest significant difference, Duncan's Multiple Range Test, and the Waller-Duncan Bayesian *k*-ratio *t* test. Particular emphasis is placed on demonstrating that the concept of comparisonwise error rate is considerably more logical, sound, and useful in pairwise multiple comparisons than the concept of experimentwise error rate. As a consequence, despite what researchers may have read in the statistical literature or what they may have heard from statistical experts, the least significant difference is appropriate whenever a pairwise multiple comparisons procedure is in order.

Additional index words: Duncan's multiple rnage test, Least significant difference, Multiple comparisons, Statistical analysis, Tukey's *w* procedure, Waller-Duncan Bayesian *k*-ratio *t* test.

PROLOGUE

Baby Bear enjoyed porridge. He had a remarkably discriminating taste for porridge and was thus very adept at recognizing outstanding porridge. Baby Bear enjoyed porridge so much that he had become a plant breeder in order to breed and develop cultivars with unique and excellent porridge-producing properties.

THE EXPERIMENT

When Baby Bear was still a young plant breeder, he had had little experience in designing experiments. One year he decided to compare the porridge yields of 15 cultivars; he wanted to compare the yield of each cultivar to the yield of each of the other 14 cultivars. That is, he wanted to make the 105 pairwise comparisons among the cultivars. Since he was naive and inexperienced in the subject of experimental design, Baby Bear conducted 105 trials at the Academic Research Farm (ARF). Each trial consisted of four replications of one of the pairs of cultivars in a randomized complete block arrangement of the eight plots. Thus the error term for a single trial had three degrees of freedom (df).

The data from each trial were subjected to analysis of variance, and the least significant difference (L.S.D.) for comparing the two cultivar means was calculated. Baby Bear used the 0.05 level of significnace; he knew this meant he should expect to falsely declare the two cultivar means different in 5% of the comparisons for which the pair of cultivars were genetically alike. Since 105(0.05) = 5.25, Baby Bear figured he would make about five Type I errors out of the 105 comparisons if all 15 cultivars were genetically alike. However, Baby Bear considered it to be rather unlikely that all 15 cultivars were genetically alike, so he did not worry a great deal about possible Type I errors. But Baby Bear did want to be able to detect differences in porridge yield if there were any.

Now it had come to pass that, while Baby Bear was harvesting his 840 (four replicates × two cultivars × 105 trials = 840) plots, the Academic Research Farm Superintendent (ARFS) had stopped by and told Baby Bear that such an excessive number of plots, just to compare 15 cultivars, would not be allowed the next year. This announcement so distressed and disturbed Baby Bear that he made an appointment to discuss his problem with Goldilocks. Goldilocks was a Statistical Lady of Great Beauty and Charm, or, as W. L. Smith, professor of statistics, Univ. of North Carolina might put it, a SLOGBAC. She had traveled extensively through many areas of the Wonderful World of Statistical Theory, but had never established a permanent residence there. Like R. A. Fisher, she knew that an understanding of statistical theory by itself did not enable one to handle statistical problems in the real world (Box, 1978, p. 270–271).

THE FIRST VISIT

After listening to Baby Bear describe his problem, Goldilocks had some advice to offer. Goldilocks suggested that Baby Bear conduct four replications of a randomized complete block design with the 15 cultivars as treatments. The experiment would occupy only 60 plots (and thus make the Academic Research Farm Superintendent happy), and would have 42 df for experimental error (and thus make statisticians happy). Baby Bear was no dumb bunny, and the advantages of this approach were immediately obvious to him. He warmly thanked Goldilocks for pointing out this method of conserving experimental resources while still meeting the objective of comparing the members of the 105 pairs of 15 cultivars. Goldilocks appreciated Baby Bear's expression of thanks and remarked that, because the L.S.D. has a comparisonwise Type I error rate and because there are 105 pairwise comparisons among the 15 cultivars, Baby Bear should expect to make 105(0.05) = 5.25, or about five, Type I errors if all 15 cultivars were exactly alike. Goldilocks pointed out that Baby Bear should not be at all surprised if he made at least one Type I error in such a situation, but to help insure against the possibility of 15 genetically alike cultivars Baby Bear might compute the L.S.D. only if the analysis of variance F test for treatment (cultivar) effects was significant at the 0.05 level. As Baby Bear prepared to leave Goldilocks mentioned that the L.S.D. is a more powerful procedure for detecting differences among treatments than other commonly employed multiple comparison procedures. While strolling back to his office Baby Bear mulled over Goldlilocks' comments, decided that Goldilocks' advice was quite reasonable, and decided to follow it for his next experiment.

[1] Contribution from the Dep. of Agronomy, Univ. of Illinois, Urbana, IL 61801. This research was partially supported by the Illinois Agric. Exp. Stn. Received 22 Apr. 1981.

[2] Professor of biometry and professor of biometry and soil fertility, Dep. of Agronomy, Univ. of Illinois, Urbana, IL 61801.

THE REVIEWERS

It too came to pass, eventually, that Baby Bear wrote a paper concerning his randomized complete block design with the 15 cultivars. After considerable amounts of rewriting, revising, editing, and polishing, Baby Bear submitted his manuscript to a Well-known, Respected, Prestigious (WKRP) scientific journal. Baby Bear's manuscript was reviewed by two peer scientists.

Reviewer number one was Papa Bear. Papa Bear was old enough to be Baby Bear's father and thus was a more mature and more experienced plant breeder. Papa Bear considered himself to be rather knowledgeable in the area of statistics; after all, he had taken a course on the design and analysis of experiments back in the early 1960s when he, himself, had been a graduate student. Papa Bear stated quite emphatically that the paper by Baby Bear was definitely in need of major revision because, according to Steel and Torrie (1960, p. 107), use of the L.S.D. was inappropriate for all possible paired comparisons. Papa Bear added that the L.S.D. was improper because the experimentwise Type I error rate for 15 treatments would be about 78% [based on probabilities of the studentized range tabulated by Harter et al. (1959)]. That is, the probability of finding at least one significant difference, even if there were no real differences among the 15 cultivars, was about 0.78. Papa Bear suggested in rather strong terms that Baby Bear might have used Tukey's w procedure, which is sometimes referred to as the honest significant difference (HSD) and has an experimentwise error rate of only 5%. Furthermore, Papa Bear added that, according to Steel and Torrie (1980, Chapter 8), this experimentwise error rate would apply to the family of all 105 pairwise comparisons. Papa Bear's real preference, however, was for use of Duncan's Multiple Range Test (DMRT) which is also described by Steel and Torrie (1980, Chapter 8). For the family of 105 pairwise comparisons the DMRT has an experimentwise error rate intermediate between that for the L.S.D. and the HSD. Papa Bear said that he preferred to use DMRT because it was common procedure to use it in articles published in this Well-known, Respected, Prestigious (WKRP) scientific journal.

Another plant breeder of some renown, Mama Bear, was reviewer number 2. She also objected to Baby Bear's use of the L.S.D. Mama Bear said that Baby Bear should have used the more modern and up-to-date Waller and Duncan (1969) Bayesian k-ratio t test (referred to as the Bayes L.S.D. by some bears). As Mama Bear pointed out, a significance level is not selected when the Waller-Duncan test is used; instead the relative seriousness of Type I and Type II errors is considered and the analysis of variance F value has a direct bearing on the magnitude of the critical value. As Mama Bear put it, "The bigger the F value, the smaller the Bayes L.S.D. critical value."

THE SECOND VISIT

Baby Bear took Papa Bear's and Mama Bear's comments to Goldilocks. Goldilocks immediately made the sage observation that there is considerable disagreement and confusion among statisticians, as well as researchers, on the subject of pairwise multiple comparisons. Goldilocks then said that, if the individual comparisons within pairs of cultivars were the conceptual units of interest to Baby Bear, then Baby Bear's use of the restricted or protected L.S.D. was indeed appropriate, and, in Goldilocks' opinion, was the procedure to choose. Goldilocks suggested that Baby Bear might read the two papers by Carmer and Swanson (1971, 1973). In addition, Goldilocks remarked that, if the entire experiment or, more specifically, the family of 105 pairwise comparisons was *not* the conceptual unit of interest, but was used instead as an efficient tool of statistical design (to reduce the number of plots from 840 to 60, while also providing an estimate of experimental error with a reasonable number of degrees of freedom), then the concept of experimentwise or familywise error rate was of importance only in the Wonderful World of Statistical Theory and should *not* be applied to real world problems where the individual comparisons are the units of concern and importance. Goldilocks said that it made little or no sense to penalize a researcher for using an efficient experimental design; the penalty of experimentwise error rate should not be inflicted upon Baby Bear just because he used a randomized complete block design with 60 plots (15 cultivars × four replications) rather than 105 trials occupying 840 plots.

Goldilocks then told Baby Bear about a computer simulation conducted by Carmer (1980, personal communication). There were 5,000 repetitions of a randomized complete block experiment with four replications of 15 treatments with identical true means. The 5% level L.S.D. was computed for each repetition; the comparisonwise Type I error rate was 0.0497 ± 0.0007 and the experimentwise Type I error rate was 0.7804 ± 0.0055. Also simulated were 285 repetitions of 105 small trials each with four replications of two treatments with identical true means. Here the use of the 5% level L.S.D. produced a comparisonwise error rate of 0.0515 ± 0.0013 which, of course, equalled the experimentwise error rate because each trial contained only two treatments. However, if a set of 105 trials is considered to be "an experiment", it may be noted that one or more Type I errors were made in 283 out of the 285 repetitions; thus the "experimentwise" error rate was 283/285 = 0.9930. This value compares favorably with the expected value, $0.9954 = [1 - (0.95)^{105}]$, for 105 independent comparisons.

Baby Bear said that it sure seemed to him that theoretical statisticians were looking for honey up the wrong tree when they invented experimentwise error rates.

Goldilocks then said that if Baby Bear wanted to reduce the probability of Type I errors he should consider changing his significance level from $\alpha = 0.05$ to either $\alpha = 0.025$ or $\alpha = 0.01$; but, since Baby Bear did not breed porridge in the Wonderful World of Statistical Theory, it was quite unrealistic to entertain thoughts of Tukey's w procedure. It was Goldilocks' opinion that DMRT should be used only by those researchers stranded somewhere between reality and the Wonderful World of Statistical Theory. Further-

more, from a practical point of view, DMRT has several drawbacks. It requires multi-valued critical values, so that the difference between cultivars required for significance depends on the number of cultivars in the experiment; it is less powerful, that is less able to detect differences than the L.S.D. Goldilocks said that, while DMRT might make some sense in the Wonderful World of Statistical Theory, it did not make much sense in the real world to think that the true difference between any two cultivars depended in any way on what other cultivars were included in the experiment.

Mama Bear's comments concerning the Waller-Duncan Bayesian *k*-ratio *t* test caused Goldilocks to pause in a thoughtful manner. Finally, Goldilocks said that Duncan and his student, Waller, had done a great deal to clarify the multiple comparisons problem and that their approach, which uses the relative seriousness of Type I and Type II errors, has considerable merit. Goldilocks added that Carmer (1976) used the idea of relative seriousness in choosing the optimal significance level for the L.S.D. and that even Duncan (1970) had stated that the Waller-Duncan procedure has a more sound logical foundation than DMRT. Goldilocks told Baby Bear that the Waller-Duncan procedure is not as simple as the restricted L.S.D.; it requires more extensive tables which are not yet as readily available to researchers as ordinary Student *t* tables. Goldilocks also pointed out that the Waller-Duncan procedure has a greater dependence than the restricted L.S.D. on the calculated *F* value, and consequently an overestimate of *F* could lead to an excessive number of Type I errors. On the other hand, as Goldilocks observed, an underestimate of *F* could lead to an excess of Type II errors and a reduction in power.

All in all it was Goldilocks' opinion that due to its simplicity and its basis on the conceptual unit of interest, the individual comparison, the restricted L.S.D. is the pairwise multiple comparison procedure of choice.

EPILOGUE

Papa Bear gruffly sputtered that his generation had been brought up on DMRT, had used it in many, many scientific papers and articles, that therefore its use must be correct, and that it was utterly unthinkable to discontinue its use.

Mama Bear answered, with only mild irritation, that her generation had abused and misused DMRT. For example, it had been applied to experiments where the treatments were quantitative levels of a factor such as fertilizer rates, seeding rates, row-spacings, or dates of planting. Mama Bear suggested that researchers might find papers by Carmer (1978), Chew (1976, 1977), Little (1978), and Petersen (1977) of interest and help. Mama Bear reiterated her position that the future lies with the Waller-Duncan Bayesian *k*-ratio *t* test, and that the sooner researchers switched to it, the better. She quoted Max Planck (1968, p. 33–34) who said, "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it."

With a smile Goldilocks said that she fervently hoped the day would come when no one used DMRT for anything, and that everyone would KISS and make up. Softly she added that KISS is an acronym for Keep It Simple, Statisticians.

And finally, Baby Bear recalled the philosophy of Carmer et al. (1979) on the role of statisticians in research. He decided however, that, since it was Goldilocks' job to provide service with a smile, it really wasn't necessary to acknowledge her contributions to his research. The main thing, he thought to himself, was to keep on eating porridge.

LITERATURE CITED

Box, J. F. 1978. R. A. Fisher, The Life of a scientist. Wiley and Sons, New York, N. Y.

Carmer, S. G. 1976. Optimal significance levels for application of the least significant difference in crop performance trials. Crop Sci. 16:95–99.

———. 1978. Use of MATRIX for single degree of freedom contrasts in factorial experiments. p. 284–287. *In* R. H. Strand (ed.) Proc. 3rd Annual Conference of the SAS Users Group International. SAS Institute Inc. Raleigh, N. C.

———, and M. R. Swanson. 1971. Detection of differences between means: A Monte Carlo study of five pairwise multiple comparison procedures. Agron. J. 63:940–945.

———, and———. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J. Am. Stat. Assoc. 68:66–74.

———, W. M. Walker, and R. D. Seif. 1979. Statisticians as reviewers. HortScience 14:9.

Chew, V. 1976. Comparing treatment means: A compendium. HortScience 11:348–357.

———. 1977. Comparisons among treatment means in an analysis of variance. USDA, ARS/H/6.

Duncan, D. B. 1970. Multiple comparison methods for comparing regression coefficients. Biometrics 26:141–143.

Harter, H. L., 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 Rep. 58-484, Vol. II.

Little, T. M. 1978. If Galileo published in Hortscience. HortScience 13:504–506.

Petersen, R. G. 1977. Use and misuse of multiple comparison procedures. Agron. J. 69:205–208.

Planck, M. 1968. Scientific autobiography and other papers. Greenwood Press, New York.

Steel, R. G. D., and J. H. Torrie. 1960. Principles and procedures of statistics. 1st Edition. McGraw-Hill, New York.

———, and———. 1980. Principles and procedures of statistics. 2nd Edition. McGraw-Hill, New York.

Waller, R. A., and D. B. Duncan. 1969. A Bayes rule for the symmetric multiple comparisons problem. J. Am. Stat. Assoc. 64:1484-1503.

13.10—Tests of homogeneity of variance. From time to time we have questioned whether two or more mean squares differ significantly. With two mean squares, a two-tailed F test is made as follows. Place the larger of s_1^2, s_2^2 in the numerator of F and read table A 14, part II, at the 2.5% level to obtain a 5% test. With a one-tailed test having $H_A: \sigma_1^2 > \sigma_2^2$, place s_1^2 in the numerator and read F at the usual 5% level; of course if $s_1^2 \leq s_2^2$ in this case, we accept H_0.

With $a > 2$ independent estimates of variance s_i^2, Bartlett (9) has provided a test. If the s_i^2 all have the same number of degrees of freedom ν, the test criterion, using logarithms to base e, is

$$M = \nu(a \ln \bar{s}^2 - \Sigma \ln s_i^2) \qquad (\bar{s}^2 = \Sigma s_i^2/a) \tag{13.10.1}$$

On the null hypothesis that each s_i^2 is an estimate of the same σ^2, the quantity M/C is distributed approximately as χ^2 with $(a - 1)$ df, where $C = 1 + (a + 1)/(3a\nu)$.

In table 13.10.1 this test is applied to the variances of grams of fat absorbed with four types of fat in the doughnut example of table 12.2.1. Here $a = 4$, $\nu = 5$. The value of M is 1.88, clearly not significant with 3 df. To illustrate the method, however, $\chi^2 = M/C = 1.74$ has also been computed.

When the degrees of freedom ν_i differ, as with samples of unequal sizes,

$$M = (\Sigma \nu_i) \ln \bar{s}^2 - \Sigma \nu_i \ln s_i^2 \qquad (\bar{s}^2 = \Sigma \nu_i s_i^2/\Sigma \nu_i)$$
$$C = 1 + \{1/[3(a-1)]\}(\Sigma 1/\nu_i - 1/\Sigma \nu_i)$$
$$\chi^2 = M/C \qquad (a-1)\text{df}$$

In table 13.10.2 this test is applied to the variances of the birth weights of five litters of pigs. Since $\bar{s}^2$ is the pooled variance (weighting by degrees of freedom), we have included a column of sums of squares as well as a column of reciprocals for finding C. The computations give $\chi^2 = 16.99$ with 4 df, showing that intralitter variances differ in these data.

The χ^2 approximation is less satisfactory if most of the df ν_i are less than 5. Special tables for this are given in the *Biometrika Tables* (10). This reference also gives the significance levels of $s_{\max}^2/s_{\min}^2$, the ratio of the largest to the smallest of the a variances. This ratio provides a quick test of homogeneity of variances that will often settle the issue, though usually less sensitive than Bartlett's test.

Unfortunately, Bartlett's test and the preceding test give too many significant results with observations that come from long-tailed distributions—distributions with positive kurtosis. An approximate test that is much less sensitive to nonnormality in the data has been given by Levene (11).

TABLE 13.10.1
BARTLETT'S TEST WHEN ALL ESTIMATES HAVE $\nu = 5$ DF

Fat	s_i^2	$\ln s_i^2$	
1	178	5.182	$M = (5)[4(4.614) - 18.081]$
2	60	4.094	$= 1.88$ (df = 3)
3	98	4.585	$C = 1 + (a + 1)/(3a\nu)$
4	68	4.220	$= 1 + 5/[(3)(4)(5)] = 1.083$
Total	404	18.081	

$\bar{s}^2 = 100.9$ $\quad \ln \bar{s}^2 = 4.614$ $\quad \chi^2 = 1.88/1.083 = 1.74$ $\quad P > 0.5$

TABLE 13.10.2
BARTLETT'S TEST FOR HOMOGENEITY OF VARIANCE WHEN s_i^2 HAVE DIFFERING DF ν_i

Litter	Sum of Squares $\nu_i s_i^2$	df ν_i	Mean Squares s_i^2	$\ln s_i^2$	Reciprocals $1/\nu_i$
1	8.18	9	0.909	−0.095	0.1111
2	3.48	7	.497	−0.699	.1429
3	0.68	9	.076	−2.577	.1111
4	0.72	7	.103	−2.273	.1429
5	0.73	5	0.146	−1.924	0.2000
$a = 5$	13.79	37			0.7080

$\bar{s}^2 = \Sigma \nu_i s_i^2/\Sigma \nu_i = 13.79/37 = 0.3727$
$(\Sigma \nu_i) \ln \bar{s}^2 = (37)(-0.9870) = -36.519$
$M = (\Sigma \nu_i) \ln \bar{s}^2 - \Sigma \nu_i \ln s_i^2 = -36.519 - (-54.472) = 17.96$
$C = 1 + [1/(3)(4)](0.7080 - 0.0270) = 1.057$
$\chi^2 = M/C = 17.96/1.057 = 16.99$ $\quad$ (df = 4) $\quad P < 0.01$

13.11—Levene's test of homogeneity of variance. As a measure of the variation within a class, Levene's test uses the average of the absolute deviations $S|X_{ij} - \bar{X}_{i.}|/n$ instead of the mean square of the deviations $s_i^2 = \Sigma(X_{ij} - \bar{X}_{i.})^2/(n-1)$. This avoidance of squaring makes the test criterion much less sensitive to long-tailed distributions. As an example, four independent samples with $n = 7$ were drawn from the t distribution with 3 df—a symmetrical long-tailed distribution—with the number 7 added to all observations to avoid negative observations. In this example, of course, we know that $H_0: \sigma_i^2 = \sigma^2$ is correct.

Table 13.11.1 shows the original data on the left and the absolute deviations $|X_{ij} - \bar{X}_{i.}|$ on the right. An observation in the data that catches the eye is the

TABLE 13.11.1
EXAMPLE OF LEVENE'S TEST OF HOMOGENEITY OF VARIANCE

	Data for Class				Absolute Deviations from Class Mean			
	1	2	3	4	1	2	3	4
	7.40	8.84	8.09	7.55	0.54	2.08	1.89	0.71
	6.18	6.69	7.96	5.65	0.68	0.07	1.76	1.19
	6.86	7.12	5.31	6.92	0.00	0.36	0.89	0.08
	7.76	7.42	7.39	6.50	0.90	0.66	1.19	0.34
	6.39	6.83	0.51	5.46	0.47	0.07	5.69	1.38
	5.95	5.06	7.84	7.40	0.91	1.70	1.64	0.56
	7.48	5.35	6.28	8.37	0.62	1.40	0.08	1.53
Total	48.02	47.31	43.38	47.85	4.12	6.34	13.14	5.79
Mean	6.86	6.76	6.20	6.84	0.589	0.906	1.877	0.827
s_i^2	0.500	1.630	7.325	1.100	0.095	0.668	3.214	0.302

TABLE 13.11.2
ANALYSIS OF VARIANCE OF MEAN DEVIATIONS

Source	df	Sum of Squares	Mean Squares	F
Between classes	3	6.773	2.258	2.11
Within classes	24	25.674	1.070	

value 0.51, which looks like a gross error, in class 3. Actually, 0.51 is not particularly unusual as the most extreme value for this long-tailed distribution.

Bartlett's test gives $\chi^2 = 11.22$, $P = 0.01$, erroneously rejecting the null hypothesis (example 13.11.1). For Levene's test we perform an analysis of variance of the mean deviations in the right half of table 13.11.1. The class means, 0.589, 0.906, etc., are our estimates of the variability within the classes. Table 13.11.2 gives the analysis.

The F value, 2.11, indicates $P > 0.10$ with 3 and 24 df—not significant. The test is approximate because the absolute deviations are not normal, and the within-class s_i^2 suggest a much higher variance within class 3 than within other classes. In fact, Satterthwaite's rule, section 12.10, suggests that the within-class mean square should have 10 df, not 24.

EXAMPLE 13.11.1—Apply Bartlett's test to the within-class s_i^2 from the original data on the left half of table 13.11.1. Ans. $M = 11.995$, $C = 1.069$, $\chi^2 = 11.22$ (3 df), P about 0.01.

EXAMPLE 13.11.2—In the data on state expenditures per pupil in 1977, the within-class mean squares s_i^2 and the degrees of freedom in five regions of the United States were as follows: Northeast, $s_1^2 = 0.1240$, $\nu_1 = 9$; Southeast, 0.0335, 6; South Central, 0.0057, 8; North Central, 0.0448, 10; Mountain Pacific, 0.0404, 10. Apply Bartlett's test. Ans. $\chi^2 = 15.35$, df = 4, $P = 0.01$.

TECHNICAL TERMS

components of variance
hierarchal classification
homogeneity of variance
intraclass correlation
mixed effects
model II
nested classification
noncentral F distribution
random effects
three-stage sampling

REFERENCES

1. Tang, P. C. *Stat. Res. Mem.* 2 (1938):126.
2. Pearson, E. S., and Hartley, H. O. *Biometrika* 38 (1951):112.
3. South. Coop. Ser. Bull. 10. 1951.
4. King, A. J., and McCarty, D. E. *J. Mark.* 6 (1941):462.
5. Newman, H. H.; Freeman, F. N.; and Holzinger, K. J. 1937. *Twins.* Univ. of Chicago Press.
6. Winsor, C. P., and Clarke, G. L. *J. Mar. Res.* 3 (1940):1.
7. Boardman, T. J. *Biometrics* 30 (1974):251.
8. Wald, A. *Ann. Math. Stat.* 11 (1940):96.
9. Bartlett, M. S. *J. R. Stat. Soc.* 4 (1937):137.
10. Pearson, E. S., and Hartley, H. O. 1954. *Biometrika Tables for Statisticians*, vol. I. Cambridge Univ. Press, Tables 31 and 32.
11. Levene, H. 1960. In *Contributions to Probability and Statistics.* Stanford Univ. Press, Stanford, Calif., p. 278.

14

Two-way classifications

14.1—Introduction. When planning a controlled experiment, the experimenter often acquires the ability to predict roughly the behavior of the experimental material. In identical environments young male rats are known to gain weight faster than young female rats. In a machine that subjects five different pieces of cloth to simulated wearing, experience shows that the cloths placed in positions 4 and 5 will receive less abrasion than those in the other positions. Such knowledge can be used to increase the accuracy of an experiment. If a treatments are to be compared, experimental units are first arranged in groups of a. Units assigned to the same group should be as similar in responsiveness as possible. Each treatment is then allocated by randomization to one unit in each group. This produces a two-way classification, since any observation is classified by the treatment it receives and the group to which it belongs.

The name given to the group varies with the type of application. In agricultural field experiments, long experience has shown that plots near one another tend to give similar yields. The group will therefore often be a compact piece of land, called a *block*. The experimental plan is described as *randomized blocks*. Another name used for the group is *replication*—meaning a single trial or repetition of the comparison between the treatments. Many experiments on human subjects show considerable variation from one subject to another. Sometimes it is possible to give each treatment to every subject on different occasions—as when comparing different analgesics for the relief of chronic headaches or different rewards for performance in a repetitive task. The objective is to make the comparisons among treatments more accurate, since they are made within subjects. The groups would then probably be called "subjects." In the abrasion tests just mentioned, the groups would be "positions," all the pieces of cloth (treatments) being tested in each position. The name used often describes the classification employed in the grouping.

Two-way classifications are frequent in surveys and observational studies, also. We encounter an example in section 11.10 in which farms are classified by soil type and owner-tenant status. In a survey of family expenditures on food, classification of the results by size of family and income level is obviously relevant.

We first present an example to familiarize you with the standard computa-

Checklist for Troubleshooting Non-estimatable Contrasts

1. Does the number of coefficients agree with the number of levels being tested for the main effect? (e.g., contrast 'Y vs X' trt 0 0 0 1 0 0 0 -1 must have eight treatment levels being accessed within your data set.

2. Do the coefficients agree with the way the data is read by the computer? A. character strings follow alphabetic sequence.
   B. numeral trt assignments follow in numerical order. (exception. ex., rates of 0 2.5 5 and 8 will be read by the computer in the order 0 5 8 2.5)

3. Does the variable before the coefficients agree with the way the coefficients were obtained? (e.g., With interactions such as placement*rate, were the placement coefficients multiplied by the rate coefficients?)

4. Does the variable in the contrast match the variable in the model statement? (e.g., With an interaction contrast such as plcmt*rate*residue, the same interaction should be in the model, not rate*plcmt*residue.)

5. Is there missing data? (i.e., Do some treatments have a different number of observations per cell?) Compare the same contrasts for different variables and see if some are estimatable for some variables while not for others.

6. If rates are present, but the methods and or sources are not included for the " 0 " rates, then the 0 rate plots must be deleted to check RATE LIN or RATE QUAD among the other rates. (i.e., the 0 rates were not found for each method and or source variable.) THE 0 RATES CAN BE INCLUDED IF THE CONTRAST IS DONE USING THE " TRT " VARIABLE.

7. In a interaction contrast, each level of the heirarchy must add up to zero whether the contrast is orthogonal or not. (e.g. contrast tillage*rate must have at least two levels of tillage used in the contrast so coefficients add to zero for that level of the heirarchy and at least two levels of rate for the same reason.

8. In an interaction contrast, the order of the variables being crossed must match the order of the variables in the classes statement. (e.g., Classes rate residue plcmt; the three way interaction contrast can only be rate*residue*plcmt, not rate* plcmt*residue.)

Orthogonal Contrasts: essentially have no overlapping of SS. (will add up to the main effect error)

Non Orthogonal Contrasts: overlapping of SS.

Table F Coefficients of Orthogonal Polynomials

		X											
k	Polynomial	1	2	3	4	5	6	7	8	9	10	$\Sigma \xi_i^2$	λ
3	Linear	−1	0	1								2	1
	Quadratic	1	−2	1								6	3
	Linear	−3	−1	1	3							20	2
4	Quadratic	1	−1	−1	1							4	1
	Cubic	−1	3	−3	1							20	$\frac{10}{3}$
	Linear	−2	−1	0	1	2						10	1
5	Quadratic	2	−1	−2	−1	2						14	1
	Cubic	−1	2	0	−2	1						10	$\frac{5}{6}$
	Quartic	1	−4	6	−4	1						70	$\frac{35}{12}$
	Linear	−5	−3	−1	1	3	5					70	2
6	Quadratic	5	−1	−4	−4	−1	5					84	$\frac{3}{2}$
	Cubic	−5	7	4	−4	−7	5					180	$\frac{5}{3}$
	Quartic	1	−3	2	2	−3	1					28	$\frac{7}{12}$
	Linear	−3	−2	−1	0	1	2	3				28	1
7	Quadratic	5	0	−3	−4	−3	0	5				84	1
	Cubic	−1	1	1	0	−1	−1	1				6	$\frac{1}{6}$
	Quartic	3	−7	1	6	1	−7	3				154	$\frac{7}{12}$
	Linear	−7	−5	−3	−1	1	3	5	7			168	2
	Quadratic	7	1	−3	−5	−5	−3	1	7			168	1
8	Cubic	−7	5	7	3	−3	−7	−5	7			264	$\frac{7}{3}$
	Quartic	7	−13	−3	9	9	−3	−13	7			616	$\frac{7}{12}$
	Quintic	−7	23	−17	−15	15	17	−23	7			2184	$\frac{7}{10}$
	Linear	−4	−3	−2	−1	0	1	2	3	4		60	1
	Quadratic	28	7	−8	−17	−20	−17	−8	7	28		2772	3
9	Cubic	−14	7	13	9	0	−9	−13	−7	14		990	$\frac{5}{6}$
	Quartic	14	−21	−11	9	18	9	−11	−21	14		2002	$\frac{7}{12}$
	Quintic	−4	11	−4	−9	0	9	4	−11	4		468	$\frac{3}{20}$
	Linear	−9	−7	−5	−3	−1	1	3	5	7	9	330	2
	Quadratic	6	2	−1	−3	−4	−4	−3	−1	2	6	132	$\frac{1}{2}$
10	Cubic	−42	14	35	31	12	−12	−31	−35	−14	42	8580	$\frac{5}{3}$
	Quartic	18	−22	−17	3	18	18	3	−17	−22	18	2860	$\frac{5}{12}$
	Quintic	−6	14	−1	−11	−6	6	11	1	−14	6	780	$\frac{1}{10}$

Experiment: Influence of Nitrogen Rate and Mowing Height on Sensor Based Detection of Nutrient Stress

Treatment	N rate lb N/1000 ft^2/month	Mowing Height inches
1	0	0.5
2	0.5	0.5
3	1.0	0.5
4	1.5	0.5
5	0	1.5
6	0.5	1.5
7	1.0	1.5
8	1.5	1.5

Replications: 4
Experimental design: CRD

CRD		**CRD**		**RCBD**	
Source of variation	df	Source of variation	df	Source of variation	df
Total (4*8)-1	31	Total (4*8)-1	31	Total (4*8)-1	31
				block	3
height	1	treatment	7	treatment	7
nrate	3				
nrate*height	3				
error	24	error	24	error	21

```
proc glm;
classes height nrate;
model yield = nrate height nrate*height;
contrast 'Nrate_lin' nrate       -3 -1 1 3;
contrast 'Nrate_quad' nrate    1 -1 -1 1;
contrast 'Nrate_cub' nrate      -1 3 -3 1;

contrast 'height*nrate_lin' height*nrate  -3 -1 1 3 3 1 -1 -3;
contrast 'height*nrate_quad' height*nrate 1 -1 -1 1 -1 1 1 -1;

means nrate height nrate*height;
run;
```

height

0.5	1.5

nrate

0	0.5	1.0	1.5	0	0.5	1.0	1.5

height

0.5 **1**	1.5 **-1**

nrate **(linear)**

0	0.5	1.0	1.5	0	0.5	1.0	1.5
-3	**-1**	**1**	**3**	**-3**	**-1**	**1**	**3**

interaction coefficients (height*nrate_lin)

-3	**-1**	**1**	**3**	**3**	**1**	**-1**	**-3**

Table A.3 Values of t

df	Probability of a numerically larger value of t								
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

df	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.0005
	Probability of a larger positive value of t								

SOURCE: This table is abridged from Table III of Fisher and Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Oliver and Boyd Ltd., Edinburgh, 1949, by permission of the authors and publishers.

TABLE 7.6.1

CORRELATION COEFFICIENTS AT THE 5% AND 1% LEVELS OF SIGNIFICANCE

Degrees of Freedom	5%	1%	Degrees of Freedom	5%	1%
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1,000	.062	.081

Portions of this table were taken from Table VA in "Statistical Methods for Research Workers" by permission of Professor R. A. Fisher and his publishers, Oliver and Boyd.

The test of H_0: $\rho = 0$, H_A: $\rho \neq 0$ is made at sight in table 7.6.1. Simply look along the row for $d.f. = 7$ and observe the position of the sample r relative to the tabular values. Our $r = 0.597$ is considerably less than the 5% level, 0.666, leading to the same conclusion as before. The test is made without considering the sign of r. Among the following correlations, observe particularly how conclusions are affected by both sample size and the size of r:

Number of Pairs	Degrees of Freedom	r	Conclusion About Hypothesis, $\rho = 0$
20	18	0.60	Reject at 1% level
100	98	0.21	Reject at 5% level
10	8	0.60	Not rejected
15	13	−0.50	Not rejected
500	498	−0.15	Reject at 1% level

Those who wish a sketch of the main features of statistics at first reading may well omit the remainder of this chapter, together with all of the next. Go to chapter 9 if you wish to learn more of enumeration statistics, or to chapter 10 for analysis of variance.