Supplement to the

JOURNAL OF NEMATOLOGY

VOLUME 23 OCTOBER 1991 NUMBER 4S

VIEWPOINT

Supplement to Journal of Nematology 23(4S):557–563. 1991.

Comparison of Treatment Means: A Statistical Fantasy[1]

J. D. MIHAIL[2] AND T. L. NIBLACK[2]

Over the last several decades, there has been a continuing debate between scientists and statisticians concerning appropriate techniques for elucidating the relationships among treatment means. To shed some light on this murky subject, the authors interviewed a scientist of world renown, Dr. Henri Hubris, Professor of Herpetology, Northeastern Central University. Dr. Hubris is widely revered among statisticians for his rigid adherence to the Highest Statistical Standards (HISS). We thought his examples and discussion were so illuminating that we decided to present them here, to assist other scientists in determining appropriate mean comparison procedures to use in their data analyses.

Dr. Hubris began the interview by pointing out that treatment comparisons are "a real snake pit." The selection of appropriate statistical procedures for analyzing data depends on the relationships among the various experimental treatments (treatment structure). The treatment structure is a separate consideration from experimental design (randomized complete block, split plot, etc.). To illustrate various treatment structures and appropriate analytical techniques, Dr. Hubris selected several of his recent studies published in the *Annual Snake Proceedings* (ASP). He also provided sample program statements (Endnote 1) from the SAS programs (13) he wrote to analyze the data from experiments described below.

PICK A WINNER: UNSTRUCTURED TREATMENTS

There is a type of experiment in which there is no obvious relationship among the several treatments imposed on experimental subjects. To illustrate this case, Dr. Hubris referred (with evident pride) to a recent study in which he tested the efficacy of 10 new antibiotics for the treatment of the devastating malady, snake flu (SNA-FU), which is characterized by rapid weight loss and death. The antibiotics tested were unrelated in their chemical composition; thus there was no reason to expect superior performance of a particular group of treatments. (Mean comparisons among related treatments are described in the next section.) Each antibiotic was administered to 15 sick snakes, each of which was weighed before treatment. After 10 days, the snakes were again weighed and the average weight

Received for publication 25 February 1991.

[1] Contribution from the Missouri Agricultural Experiment Station Journal Series No. 11,364.

[2] Plant Sciences Unit, Plant Pathology Program, University of Missouri, Columbia, MO 65211.

The authors gratefully acknowledge Dr. H. Hubris for taking many hours from an extremely busy schedule to share his insights. The authors similarly acknowledge the following colleagues for providing helpful reviews of this manuscript: Dr. J. English, Dr. S. Pueppke, and Dr. T. Wyllie.

TABLE 1. Results of multiple comparison procedures applied to the effects of 10 antibiotics on weight gain of snake-flu infected snakes 10 days after antibiotic treatment. Means followed by the same letter are not significantly different.

Antibiotic	Weight gain†	Multiple-comparison procedure‡			
		W-D	PLSD	SSD	HSD
K	4.44	a	a	a	a
Z	4.37	a	a	a	a
W	3.01	b	b	b	b
P	2.98	b	b	b	b
M	2.96	b	b	b	b
Q	2.87	b	b	b	b
S	1.13	c	c	c	c
Y	1.08	c	cd	cd	c
F	0.88	d	d	cd	cd
V	0.67	e	e	d	d

† Weight gain is the mean for 15 snakes.

‡ W-D = Waller-Duncan k-ratio *t*-test; PLSD = Fisher's Protected Least Significant Difference; SSD = Scheffe's Significant Difference Method; HSD = Tukey's Honestly Significant Difference. WD was computed with a k-ratio of 100, a Type I/Type II error seriousness ratio roughly correspondent to $P = 0.05$. PLSD had a comparison-wise error rate of $P = 0.05$. SSD and HSD had experiment-wise error rates of $P = 0.05$. Duncan's multiple-range test and Student-Newman-Keul's Multiple Range Test gave the same mean separations as PLSD.

gain was computed for each of the 10 treatments (Table 1). Because the objective was to select the best possible antibiotic, the analysis should compare every possible pair of antibiotics; thus a multiple mean comparison procedure is in order.

Dr. Hubris explained that the selection of the proper multiple comparison procedure is not a trivial task. Some statisticians argue that all multiple comparison procedures are an egregious breach of HISS, and indeed, many of the most heinous violations of HISS are related to the misuse of these procedures (2,6,9,12,15). The selection of the most suitable multiple comparison procedure depends on the effect of the Type I and Type II error rates (Endnote 2) relative to the objectives of an experiment.

Dr. Hubris applied six mean comparison procedures to his SNAFU antibiotic data to illustrate some of the differences among the methods (Table 1). He noted that five of these tests had been compared by Boardman and Moffitt (1) for data sets with 10 means and a Type I comparison-wise error rate of 5% (Endnote 2). When comparing all possible pairs of means, the five tests may be listed as follows, in order of decreasing magnitude of the Type I experiment-wise error rate: Fisher's Protected Least Significant Difference (LSD), Duncan's multiple-range test, Student-Newman-Kuel's Multiple Range Test, Tukey's Honestly Significant Difference, and Scheffe's Significant Difference Method. Thus, Fisher's Protected LSD and Scheffe's Method will declare the largest and smallest number of significantly different mean pairs, respectively (1,3). Fisher's Protected LSD is a least significant difference procedure requiring that the overall analysis of variance F-test be significant prior to pair-wise comparison of means. A sixth test used was the Waller-Duncan k-ratio *t*-test, which allows the researcher to specify an acceptable Type I/Type II error ratio. (Detailed discussions of various procedures are given by Chew [3].)

Fisher's Protected LSD, Duncan's multiple-range test, and Student-Newman-Kuel's Multiple Range Test all gave the same partitioning of experimental means into five groups (Table 1). The more conservative Tukey's and Scheffe's procedures split the treatment means into only four groups. The Waller-Duncan k-ratio *t*-test, with a k-ratio of 100, gave a mean separation into five distinct groups. All six procedures are available on computer statistical packages (10,13). Thus, the choice of multiple comparison procedure need not rest on computational considerations *but should be guided by the objectives of the experiment.* For example, is it critical to unearth all possible differences among means? If so, a test such as Fisher's Protected LSD is an appropriate choice (3,15). If it is essential that the fewest pairs of means be declared significantly different, then a more conservative procedure such as Scheffe's method may be used. Dr. Hubris reiterated the point made by previous researchers (9): Although the debate concerning an appropriate multiple comparison procedure could be endless, the important thing is to

know whether or not a multiple comparison procedure is appropriate at all (11).

Cluster analysis is an alternative procedure to the multiple comparison procedures outlined above. Multiple comparison procedures separate treatment means into homogeneous but often overlapping groups, whereas cluster analysis is a procedure that divides treatment means into homogeneous, nonoverlapping groups (9,14). Cluster analysis has not been used frequently, but Gates and Bilbro (5) described applications of the procedure to experiments in agronomy and provided examples of the various computational steps. Although the technique has not been widely compared with traditional multiple comparison procedures (9), the separation of means into nonoverlapping groups focuses on similarities within groups rather than on differences among them.

A Strike for Independence: Planned Orthogonal Contrasts

Unlike the experiment described above, most studies involve naturally related treatments. This structure suggests a different strategy for analysis of the data. To illustrate this type of experiment, Dr. Hubris described a follow-up to the antibiotic experiment. He designed an experiment to compare three formulations of antibiotic K (two orally administered, K_{O1} and K_{O2}; one dermally applied, K_D) and two formulations of antibiotic Z (both orally administered, Z_{O1} and Z_{O2}) with a standard antibiotic treatment (F). The experimental protocol and data collection were the same as for the first experiment.

The treatment structure of this second experiment immediately suggested several hypotheses of interest:

1) Is the effect of the standard antibiotic (F) different from the effects of the other five treatments?

2) Are the two oral formulations of antibiotic K different from the dermal formulation?

3) Are the formulations of antibiotic K different from the formulations of antibiotic Z?

4) Are the two oral formulations of antibiotic Z different?

5) Are the two oral formulations of antibiotic K different?

Dr. Hubris explained the comparisons this way: For each of these hypotheses, we really wish to compare two groups of means. Consider the first hypothesis. Group 1 is composed of the single treatment F and Group 2 comprises all five formulations of the new antibiotics. Our null hypothesis (that there is no difference between the groups) may be stated as:

$$F = (K_{O1} + K_{O2} + K_D + Z_1 + Z_2)/5,$$

or, "the mean of F is equivalent to (no different than) the mean of all the other treatments combined." Furthermore, we can restate the equation by subtracting:

$$F - (K_{O1} + K_{O2} + K_D + Z_1 + Z_2)/5 = 0.$$

To eliminate the necessity of division, multiply both sides of the equation by 5, giving:

$$5F + (-1)K_{O1} + (-1)K_{O2} + (-1)K_D + (-1)Z_1 + (-1)Z_2 = 0.$$

Because the coefficients of this equation sum to zero, this type of comparison of means is termed a "contrast" (3,6).

The coefficients for the five contrasts to test the five hypotheses listed above are given (Table 2). Notice that for each contrast, the coefficients sum to zero. A second feature of these contrasts may be illustrated by considering Contrasts 1 and 2. If their corresponding coefficients (from Table 2) are multiplied and summed for all six pairs of coefficients, the sum is zero. Thus, for Contrasts 1 and 2,

$$(5)(0) + (-1)(1) + (-1)(1) + (-1)(-2) + (-1)(0) + (-1)(0) = 0.$$

Similarly, for Contrasts 3 and 5, coefficients (from Table 2) may be multiplied and summed as follows,

$$(0)(0) + (2)(1) + (2)(-1) + (2)(0) + (-3)(0) + (-3)(0) = 0,$$

and so on. This property of the coefficients results from the independence of the contrasts (3,6,15). By comparing the remain-

TABLE 2. Coefficients used to multiply by treatment means for planned orthogonal contrasts.

Contrast†	Antibiotic formulation						
	F	K_{O1}	K_{O2}	K_D	Z_{O1}	Z_{O2}	P‡
1	5	−1	−1	−1	−1	−1	0.00
2	0	1	1	−2	0	0	0.00
3	0	2	2	2	−3	−3	0.00
4	0	0	0	0	1	−1	0.04
5	0	1	−1	0	0	0	0.27
Antibiotic mean	3.1	4.9	5.0	4.5	4.1	3.8	

† Contrasts 1–5 refer to the five hypotheses proposed in the text. Notice that the coefficients for each contrast sum to 0 and that the product of the corresponding coefficients for any pair of contrasts sum to 0. The coefficient for a treatment not included in a contrast is 0.

‡ Computed probability level for the single degree of freedom contrasts (in this case, the contrast is significant if $P < 0.05$).

ing pairs of contrasts in this way, the reader will discover that all five hypotheses are independent of each other. Such contrasts are termed "orthogonal contrasts." The name "planned orthogonal contrasts" implies (and indeed requires) that the comparisons between treatments, or groups of treatments, be planned *before* the data are gathered. The practical reason for planning is to assure that the hypotheses of interest are independent and can be tested at a desired P level before time and effort are invested in experimentation. (To be sure, it is possible to construct nonorthogonal sets of contrasts and test them at higher P levels.)

For any k means there are $(k - 1)$ possible orthogonal contrasts, each with a single degree of freedom. In this antibiotic experiment, six treatments were tested; thus, five orthogonal contrasts were possible. From the calculated P values for each contrast (Table 2), it is evident that only the contrast comparing the two oral formulations of antibiotic K was not significant ($P < 0.05$). Dr. Hubris, a devotee of efficiency, noted that only five hypotheses were tested using planned orthogonal contrasts, whereas comparison of all possible mean-pairs would have resulted in 15 different comparisons!

A Crossing of Roads: Factorial Experiments

One of the most common types of experiments in biology involves measuring a particular variable in response to two or more experimentally imposed qualitative factors. To illustrate this factorial type of experiment, Dr. Hubris proceeded to describe an experiment in which he examined the production of snake venom as influenced by diet and activity of the subjects. Two dietary regimes (mice or gerbils) and three activity regimes (sleep, two periods of daily slithering, or one daily encounter with a large snake of another species) were combined in all possible combinations to give a total of six treatments. Five snakes were randomly assigned to each of the six treatments. After one week of treatment regimen, venom was extracted using the most modern instrument, VIPER (Venom Induction-Purification-ExtractoR).

The appropriate analytical technique for this experiment (and for factorial experiments in general) is the analysis of variance (3,6,8,15,16). Analysis of variance involves partitioning the variability in the data (as measured by the mean square) into portions attributable to the various experimental factors (main effects), interactions among these factors, and undetermined sources (error, or residual). Partitioning of the mean square can become complex, and assistance may be derived from examples provided elsewhere (8). Dr. Hubris provided a digression on theoretical assumptions of analysis of variance and transformation of data (Endnote 3).

For the venom-production experiment (Table 3), Bartlett's test for homogeneity

of variance (6,15) indicated that this assumption was sufficiently well met to eliminate the need for transformation of the data. The assumption of additivity was also satisfied, because the interaction mean square was not significant ($P = 0.887$). The assumption of error independence was met through randomization of experimental units, and the interpretation of the analysis could proceed without further consideration of the assumptions underlying the analysis of variance.

From the analysis of variance, it was apparent that the effect of diet on venom production was not significant ($P = 0.611$). Thus, all the information concerning the effects of activity are contained in the column means. Although it would be inappropriate to apply a multiple comparison procedure (e.g., Fisher's Protected LSD) to all possible pairs of the six treatment means, it is quite appropriate to use such a procedure to compare the differences among the three column means representing the mean effects of sleep, slithering, and encounters with larger snakes. In doing so, one of the additional advantages of analysis of variance becomes clear. The column means contain 10 observations, whereas the individual treatment means contain only five observations. Thus, there is hidden replication in the factorial design.

If a significant interaction between diet and activity had been found, it would have been necessary to take a different approach to the separation of treatment means. In this case, Fisher's Protected LSD (or other multiple-comparison procedure) could have been applied twice; first to the three means within the mouse diet and then to the three means within the gerbil diet.

TABLE 3. Analysis of variance of the factorial experiment measuring the production of snake venom as related to diet and activity.

	Venom (ml) produced per snake			
Diet	Sleep	Slithering	Close encounters	Diet mean
Mice	1.88	2.44	4.52	2.95
Gerbils	1.92	2.26	4.30	2.83
Activity mean	1.90	2.35	4.41	

Source of variability	df†	Mean square	Probability
Main effects			
Diet	1	0.108	0.61
Activity	2	17.910	0.00
Interaction	2	0.049	0.89
Residual	24	0.406	

† df = degrees of freedom.

CONNECT THE DOTS: RESPONSE TO QUANTITATIVE TREATMENTS

The previous experiment considered the response (venom production) to two qualitative factors. However, it is often desirable to examine responses to graded levels of quantitative factors. Dr. Hubris pointed out two of his recent studies that fell into this category: 1) the effect of five levels (0,10, 20, 40, and 100 mg/kg) of the growth hormone FANG (Farley's Adolescent Nerve Generator) on the weight gain of adolescent snakes and 2) the effect of temperature on respiratory function as measured by the Respiratory And Total Thoracic and Lung Expansion (RATTLE) index.

For dose-response experiments such as these, a multiple-comparison procedure would be an unthinkable violation of HISS by ignoring the graded level of the treatments. The appropriate analytical tools would be regression analysis or some other curve-fitting procedure (15). These topics are well beyond the intended scope of this discussion but might form the basis for a future interview with another revered practitioner of HISS.

EPILOGUE

At this point, the conversation had been quite lengthy, and it was clear to the authors that Dr. Hubris' patients had worn thin. The discussion had provided several simple yet elegant examples of the appropriate uses of various statistical techniques. Dr. Hubris concluded that, although a scientist of his achievements rarely requires it, statistical help is never far away. Excel-

lent textbooks cover all of these subjects in pointed detail (8,15). A statistical procedure should not be selected after data have been collected. Rather, considerations of appropriate statistical analysis should be an integral part of the experimental design process. Remember: Statisticians Like Invitations To Help EaRly (SLITHER!).

Literature Cited

1. Boardman, T. J., and D. R. Moffitt. 1971. Graphical Monte Carlo Type I error rates for multiple comparison procedures. Biometrics 27:738–744.

2. Carmer, S. G., and W. M. Walker. 1982. Baby bear's dilemma: A statistical tale. Agronomy Journal 74:122–124.

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

4. Finney, D. J. 1989. Was this your statistics textbook? V. Transformation of data. Experimental Agriculture 25:165–175.

5. Gates, C. E., and J. D. Bilbro. 1978. Illustration of a cluster analysis method for mean separation. Agronomy Journal 70:462–465.

6. Gilligan, C. A. 1986. Use and misuse of the analysis of variance in plant pathology. Pp. 225–261 *in* D. S. Ingram and P. H. Williams, eds. Advances in plant pathology, vol. 5. New York: Academic Press.

7. Krajewski, P. 1990. Heterogeneity of variance in field experiments: Some causes and practical implications. Journal of Agricultural Science 115:83–93.

8. Little, T. M., and F. J. Hills. 1978. Agricultural experimentation: Design and analysis. New York: John Wiley and Sons, Inc.

9. Madden, L. V., J. K. Knoke, and R. Louie. 1982. Considerations for the use of multiple comparison procedures in phytopathological investigations. Phytopathology 72:1,015–1,017.

10. Norusis, M. J. 1990. SPSS/PC+, 4.0 base manual. Chicago: SPSS, Inc.

11. Perry, J. N. 1986. Multiple-comparison procedures: A dissenting view. Journal of Economic Entomology 79:1,149–1,155.

12. Petersen, R. G. 1977. Use and misuse of multiple comparison procedures. Agronomy Journal 69: 205–208.

13. SAS Institute, Inc. 1989. SAS/STAT User's Guide, version 6, fourth edition, volumes 1 and 2. Cary, NC: SAS Institute, Inc.

14. Scott, A. J., and M. Knott. 1974. A cluster analysis method for grouping means in the analysis of variance. Biometrics 30:507–512.

15. Sokal, R. R., and F. J. Rohlf. 1981. Biometry, 2nd ed. San Francisco: W.H. Freeman and Co.

16. Swallow, W. H. 1984. Those overworked and oft-misused mean separation procedures—Duncan's, LSD, etc. Plant Disease 68:919–921.

Endnotes

Endnote 1: Sample SAS code. The program code listed below is derived from the SAS programs Dr. Hubris used to analyze data from the experiments described. DATA ONE refers to the experiment described under the section heading "Pick a Winner: Unstructured Treatments"; DATA TWO refers to "A Strike for Independence: Planned Orthogonal Contrasts"; and DATA THREE refers to "A Crossing of Roads: Factorial Experiments." These samples are included for illustrative purposes only; these program segments cannot illustrate all the intricacies of programming SAS.

```
DATA ONE;
INFILE 'A:EXPT1';
INPUT ANTIBIO SNAKE WTGAIN;
PROC ANOVA;
CLASSES ANTIBIO;
MODEL WTGAIN=ANTIBIO;
MEANS ANTIBIO/WALLER;

DATA TWO;
INFILE 'A:EXPT2';
INPUT ANTIBIO SNAKE WTGAIN;
PROC SORT; BY ANTIBIO;
PROC GLM;
CLASSES ANTIBIO;
MODEL WTGAIN=ANTIBIO;
CONTRAST 'F VS. OTHERS' ANTIBIO 5 -1 -1
  -1 -1 -1;
CONTRAST 'ORAL V. DERMAL K' ANTIBIO 0
  1 1 -2 0 0;
CONTRAST 'K VS. Z' ANTIBIO 0 2 2 2 -3 -3;
CONTRAST 'Z1 VS. Z2' ANTIBIO 0 0 0 0 1 -1;
CONTRAST 'K1 VS. K2' ANTIBIO 0 1 -1 0 0 0;

DATA THREE;
INFILE 'A:EXPT3';
INPUT DIET ACTIVITY SNAKE VENOM;
PROC GLM;
CLASSES DIET ACTIVITY;
MODEL VENOM = DIET ACTIVITY
DIET*ACTIVITY;
MEANS ACTIVITY/LSD;
```

Endnote 2: Type I and Type II error. Error, as related to statistical analysis, does not refer to errors that might occur in the measurement of experimental variables. Rather, error refers to the probability of making an incorrect inference from the analyzed data.

For the comparison of two means, a Type I error occurs if the two means are declared different, when in fact they are not. For most analyses, the Type I error rate is selected by the researcher at the 5% or 1% level ($P = 0.05$ or $P = 0.01$). As Chew (3) pointed out, there are actually two kinds of Type I error when the analysis involves the comparison of three or more means. An "experiment-wise" error rate of 0.05 would mean that 5% of the experiments will declare at least one mean-pair to be significantly different when it is not. A "comparison-wise" error rate of 0.05 implies that 5% of the mean-pairs compared within a single experiment will be declared significantly different when they are not.

Type II error is the probability that two means that are truly different will be declared the same as the result of the analysis. Clearly, Type I and Type II errors are inversely related. An experiment-wise error rate of 5% is more conservative than a comparison-wise error rate of 5% in that fewer mean-pairs will falsely be declared different. However, this conservative approach will increase the probability of failure to detect true differences (Type II error). Determining the appropriate balance between these two types of error depends on the experimental objectives (3).

Endnote 3: Assumptions and violations thereof in analysis of variance. Dr. Hubris explained (rather venomously) that introductory statistics courses often leave students with the impression that if the theoretical assumptions of analysis of variance are not rigorously met, the analysis should not even be attempted. As Gilligan (6) pointed out, however, failure to meet one or more of the assumptions jeopardizes *not* the computations, but the inferences drawn from the analysis. The primary assumption of analysis of variance (4,15) is that the "error" portion of the individual measurements must be independent of one another. This is easily achieved by randomly assigning the snakes (or other experimental units) to the various treatments. It is also necessary that there be no *a priori* interaction between the various treatments in a factorial experiment. This is often termed the assumption of additivity of treatment and block effects. The third assumption of analysis of variance is that the variances of all treatments are the same, or homogeneous. This homogeneity of variance has also been termed homoscedasticity (15), a term that caused Dr. Hubris' beady little eyes to positively gleam. The final assumption of this analysis is that the error terms are normally distributed.

Although the fulfillment of these assumptions is an admirable goal of HISS, the data from many well-designed and well-executed experiments fail to satisfy one or more of the assumptions. The assumption of the independence of the error terms is inviolable, and thus the randomization of experimental units is critical. Failure to meet the other assumptions may be addressed through transformation of the data. Selection of a proper transformation depends on the type of measurements being made. Details of these considerations are given by Finney (4) and by Little and Hills (8). Krajewski (7) treats the aspects of field experiments that may contribute to lack of variance homogeneity.

Tests of the assumption of homogeneity of variance, such as Bartlett's test, are available in statistical software (SPSS or SAS). For SAS novices, Bartlett's test can easily be computed by following the example in Sokal and Rohlf (15), using variances printed by the MEANS procedure. If the variances are not equal, an appropriate transformation should be selected and the analysis of variance repeated.