On the Status of Statistics in PHYTOPATHOLOGY

Steven B. Johnson and R. D. Berger

Graduate student and professor, Department of Plant Pathology, University of Florida, Gainesville 32611.
Florida Agricultural Experiment Station Journal Series Paper 3456.
Appreciation is expressed to J. A. Cornell, Department of Statistics, University of Florida, for critical review of the manuscript.
Accepted for publication 12 March 1982.

We have become increasingly concerned about the inappropriate use of statistical methods, especially multiple comparison procedures, in many articles appearing in PHYTOPATHOLOGY. Authors abused multiple comparison procedures in two-thirds of the nearly 200 tables and figures in PHYTOPATHOLOGY Volume 70 (October issue) through Volume 71 (September issue) in which they were used.

Plant pathologists are not the only researchers that have used multiple comparisons inappropriately. Biometricians have criticized horticulturists three times (1,2,5) and agronomists once (6).

The most common statistical abuse that regularly appears in PHYTOPATHOLOGY is the use of multiple comparison procedures to compare the responses of host plants exposed to different levels of quantitative factors such as inoculum density, temperature, soil (or medium) amendments, space, or time. Regression analysis or curve fitting is the appropriate procedure to use for these comparisons. If either of these techniques is significant, the treatment means are different and no multiple comparison procedure is warranted or even proper. For some experimental designs, other statistical procedures such as orthogonal polynomial contrasts or pairwise comparisons may be applicable.

Several articles have appeared that describe the correct use of multiple comparison procedures (1–3,5). In an excellent USDA bulletin, Chew (3) states: "In comparing the effects of, say, 10, 20, 30, and 40 p/m of a certain chemical, if the regression of the response on concentration or if any component of the sum of squares for concentrations is significant, then no multiple comparison procedure is necessary. ALL concentrations are significantly different in their effects. In fact, not only will 10 and 20 p/m be different, but so also will 10 and 10.1 p/m. The difference, of course, between the effects of 10 p/m and 10.1 p/m will be extremely small. However, the usual statistical test of significance is not concerned with the magnitude of the difference, but only whether a true difference exists, no matter how small."

In summary, *multiple comparison procedures should not be applied to the average responses at levels of a quantitative factor.*

Another statistical abuse that has appeared in PHYTOPATHOLOGY is the use of multiple comparison procedures in factorial experiments to test significance of the averaged main effects of the factors without verifying that interactions were absent. The proper approach is to partition the degrees of freedom and associated sums of squares for treatments into those attributable to effects of the single factors and to the interaction effects among them. First, the interactions, starting with those of the highest order, are tested with an appropriate error term. Then, if no significant interaction effects are found, the main effects can be tested, again with an appropriate error term. However, should an interaction appear to be significant, it would be improper to test the main effects and state inferences about the main effects of the interacting factors. Blanket application of a multiple comparison test for treatments without first testing the interactions will invariably lead to incorrect conclusions regarding the nature and magnitude of the treatment response. In summary, *multiple comparison procedures should not be applied to main-effect treatments in factorial experiments when significant interactions are present.*

DISCUSSION

To exemplify our criticism on the misuse of multiple comparison procedures, we want to discuss results typical of those that have appeared in PHYTOPATHOLOGY. We have refrained from using tables from recent issues of PHYTOPATHOLOGY to illustrate statistical abuses. It is not our intention to single out specific authors for abuses committed by many. Additionally, reviewers or editors may have suggested or requested an inappropriate multiple comparison procedure; thus, the error is not of the author's volition. In Table 1, we present some of our results for levels of a quantitative factor. In this case, the survival of sclerotia was examined after burial at different depths in soil for a period of time. If Duncan's new multiple range test is applied, inappropriately, the treatment means of sclerotial survival at different burial depths are separated into two significance groups. The comparison test would be interpreted to mean that burial at 10.0 cm would not be significantly better than burial at 5.0 cm. The grower might then choose a light tilling operation to avoid the extra energy expenditure required for deeper plowing. The proper statistical approach would be to use regression of sclerotial survival on depth. Properly analyzed, a significant linear response is present. In parallel with Chew's (3) statement above, burial at 10 cm would allow significantly less sclerotial survival than burial at 5.0 cm, and even 5.1 cm would be significantly less than 5.0 cm. The linear response may now be included in a decision process. The cost of deeper plowing could be balanced against the probable benefit of reduced sclerotial survival.

Researchers must be made aware that there are also special problems associated with the use of regression and curve fitting techniques: Outliers, patterns of residuals, and excessive scatter of the data are often encountered. Some of these problems can be corrected in future experiments by choosing different treatment levels and making improvements in technique.

Factorial experiments should be analyzed as such, a procedure not always followed by those who publish in PHYTOPATHOLOGY. Examples were found in PHYTOPATHOLOGY in which all treatments of a factorial experiment were compared with a single multiple comparison procedure. It would be an abuse of a multiple comparison procedure to use it to analyze a factorial experiment with treatment means when, say, four mycorrhizal species are being tested across three host cultivars for effects on yield. As stated above, the proper approach would be to test first for the presence of interactions. If an interaction is present, then the response of the mycorrhizal species on each host can be evaluated separately, but *not* the mean response of the mycorrhizal species. This would inappropriately combine the interacting effects and would mask the proper ranking of the responses.

In PHYTOPATHOLOGY, a problem obviously exists when authors abuse statistical procedures in >60% of the cases. What then is the solution? Authors must assume the major portion of responsibility for correct analysis of their data. Helpful advice from a statistician on the choice of correct procedures will avoid such errors. More importantly, the statistician's advice should be sought before the experiment is conducted so that the most beneficial experimental

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.

0031-949X/82/08101402/$03.00/0

TABLE 1. An example of inappropriate use of a multiple comparison procedure for burial depths (a quantitative factor)

Burial depth (cm)	Number of sclerotia recovered[y]
5	83 a[z]
10	71 a
20	37 b
30	21 b

[y] Means are the averages of four replications.
[z] Numbers followed by the same letter are not significantly different, $P = 0.05$, according to Duncan's new multiple range test.

design can be chosen. The assistance of a statistician at an early stage in designing the experiment frequently permits evaluation of relationships that the researchers may not have been able to detect or test otherwise.

The ultimate and perhaps extreme solution to avoid statistical abuses in papers published in PHYTOPATHOLOGY would be to have statisticians review the manuscripts. To do this properly, such reviewers ideally would need enough of the raw data and sufficient explanation of the experimental design to check the analysis, as suggested elsewhere (5). The publication of analysis of variance (ANOVA) tables should be encouraged when these could increase the reader's understanding of the treatment effects (see Table 2 in reference [4] for an example of ANOVA with a quantitative factor).

Reviewers, Senior Editors, and the Editor-in-Chief must share partial blame with the authors when tables or figures with inappropriate analyses are published in PHYTOPATHOLOGY. This is not to say that all persons in these positions should be authorities in the use of statistics, but they should recognize those who are and seek their professional opinions as reviewers. The recognition and elimination of the most common of the abuses noted above would improve the respectability of our journal.

LITERATURE CITED

1. Chew, V. 1976. Comparing treatment means: A compendium. HortScience 11:348-357.
2. Chew, V. 1976. Uses and abuses of Duncan's Multiple Range Test. Proc. Fla. State Hortic. Soc. 89:251-253.
3. Chew, V. 1977. Comparisons among treatment means in analysis of variance. U.S. Dep. Agric. Tech. Bull. H-6. 64 pp.
4. Dean, J. L. 1979. Sugarcane mosaic virus: Shape of the inoculum infection curve near the origin. Phytopathology 69:179-181.
5. Little, J. M. 1978. If Galileo published in HortScience. HortScience 13:504-506.
6. Petersen, R. G. 1977. Use and misuse of multiple comparison procedures. Agron. J. 69:205-208.