DOCUMENT RESUME

ED 135 851 TM 006 081

AUTHOR Games, Paul A.
TITLE Nesting, Crossing, Type IV Errors, and the Role of Statistical Models.
PUB DATE [Apr 77]
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Variance; *Error Patterns; *Mathematical Models; *Statistical Analysis; *Tests of Significance
IDENTIFIERS *Type IV Error

ABSTRACT

Games (1973) criticized the Marascuilo and Levin (1970) "nested" or simple effects design as a poor substitute for the usual logic of a factorial analysis of variance. Marascuilo and Levin's use of "nesting" is completely idiosyncratic and contrary to well-established usage, thus confusing the student. Proper usage is illustrated. Statistical usage modifies models to fit the reality of the data, just as scientific models are modified to fit the facts. All present scientific models are wrong. We seek gradual improvements so our models come ever closer to the truth, but we recognize them as approximations with a certain margin of error. Similarly in data analysis we search for a simple model that describes the data within the limits of sampling error. Texts and articles reflecting this approach are cited. Marascuilo and Levin (1976) actually claim that Type IV errors are as important as Type I and Type II errors. They advocate extreme rigidity as a virtue disguising it under the name of "elegance". The flexibility desired in a good scientist is equally desirable in a good data analyst. (Author)

Nesting, Crossing, Type IV Errors, and the Role of Statistical Models

Paul A. Games

The Pennsylvania State University

The general title of this symposium implies we shall consider all types of errors, from Type I to Type IV. I assume all the participants and listeners agree on the seriousness of Type I and Type II errors. Graduate statistics courses should include consideration of the basic trade-offs between Type I and Type II errors. Unfortunately, given everything else constant, whatever we do to reduce Type I errors decreases power and increases Type II errors. Nothing comes free; to keep alpha constant and reduce the risk of Type II errors we must increase n, reduce MS_E by improving the design, improving the reliability of the test, etc. Type III errors often have been proposed, but have never stuck. My favorite of the alternatives is from Kimball (1957): Giving the right answer to the wrong problem. Levin and Marascuilo (1972, hereafter L&M used generically for any of their writings) define a Type IV error, as Levin has presently covered. Although other parts of their papers are interesting and valuable, their statements about Type IV errors are confused and confusing, and merely create a fog of intellectual ambiguity that wastes the time and effort of otherwise sophisticated statistics users such as Austin Beggs.

I agree that the original L&M 1970 AERJ article was a worthwhile contribution otherwise. It introduced many behavioral scientists to practices such as using a single df interaction contrast when this matches the E's apriori prediction. Such contrasts had been included in earlier textbooks, J. C. R. Li's Statistical Inference (1964) and J. L. Myers Fundamentals of Experimental Design (1966) but these books were never very

popular so that L&M's presentation of this material undoubtedly reached new readers. Other parts of this article I find less useful, and three parts of this and later papers I object to. Unfortunately, the remainder of this paper will deal with the disagreements and will ignore the major areas of agreement.

I often advise students to select a single contrast that reflects their apriori experimental predictions either on main affects or interactions. Let us assume a design with, say, $A_3 \times B_4$ and hence $df_{int} = 6$. Let us assume we want to maintain the familywise Type 1 error rate, FWI, at .05 for the AB interaction. The student may run a t (or F) on a single interaction contrast at $\alpha = .025$, and also run the omnibus F on the SS(residual interaction) with df = 5 at $\alpha = .025$, thus preserving FWI = .05. What is the advantage of this two part procedure over simply doing the omnibus F on the interaction by MS_{int}/MS_E at $\alpha = .05$? As pointed out in an earlier article (Games, 1971), given = n's and = σ^2's, the omnibus interaction F is the average of six different df = 1 F's on six orthogonal interaction contrasts. The five contrasts that are orthogonal to the predicted contrast may all be zero, thus drastically lowering the mean F of the omnibus interaction $F = {}^{6}\Sigma F_j/6$. If the prediction is correct, the one single interaction tested at $\alpha = .025$ will have far greater power than will the omnibus F on 6df in such a situation. Thus we maintain control of FWI and gain power by judicious apriori planning.

What would L&M have us do on this interaction contrast? They contend that when an F is used, the only legitimate method of contrast testing is the Scheffe'. It is easy to demonstrate that using the Scheffe' the single contrast will be significant only if that contrast is so large that it alone will yield a significant omnibus F interaction test. Thus L&M insist

on throwing away the superior power that is possible with better ways of controlling FWI then the Scheffe', all in the name of avoiding Type IV errors. They reduce you to the same low power situation that is automatically inherent in the omnibus F test in such a situation. Throwing away power is one of the things I teach my students not to do; why deliberately increase Type II errors?

The second point in which I disagree with L&M is on their idiosyncratic use of the word nesting to describe incorrectly an experimental situation and corresponding statistical model. There are two basic terms that are crucial to a description of experimental designs and of statistical models. They are nested and crossed. These terms are defined as follows:

Two factors are crossed if each level of each factor appears with each level of the other factor. If each subject, S, appears with each level of T, we say S and T are crossed and represent this as S x T as in a repeated measures design. We may use subscripts on the factors to represent the number of levels. Thus $S_n \times T_5$ indicates that n subjects are crossed with 5 levels of time.

If each level of factor A occurs within only one level of factor B, then we say A is nested in B. This may be represented as A in (B). In a completely randomized one factor design where 50 subjects are randomly divided into 10 subjects under each of 5 levels of factor A, we would write S_{10} in (A_5). A completely randomized factoral design with 2 levels of A, and 3 of B may be represented as S_n in $(A_2 \times B_3)$. Since subjects are nested in the AB crossing, this shows that any one subject appears in only one of the 6 cells formed by the AB crossing. Lee (1975) has written a behaviorally oriented AOV text using crossing and nesting notations

throughout that should be consulted if an expansion of this brief explanation is desired.

All the experimental designs that Kirk (1968) covers up to page 318, and that Winer (1971) covers up to page 604 easily can be described using these two basic terms. For example the design that Kirk (1968) calls a split-plot design, and represents as a SPF-2.3 design, and that Winer (1971) calls a two factor experiment with repeated measures on one factor may be represented briefly as an $[S_n \text{ in } (A_2)] \times B_3$ design. That is, the cells of factor A and B are crossed, again giving 6 cells. However, subjects are nested under A, but crossed with B. This use of crossing and nesting is utterly basic. It was used by Bennett and Franklin (1954) and by Cornfield and Tukey (1956) for the derivation of expected MS's in factorial designs. It is used in the univariate AOV computer programs of BMD08V, BMDP2V, ANOVR, RUMMAGE, SAS, and the multivariate MANOVA program BMD12V, and probably many others that I have not encountered. Despite the essential and basic quality of this term, nested, and the unaminimity of prior usage, L&M blithly give a completely different idiosyncratic usage. They take an S_n in $(A_2 \times B_3)$ conventional factorial design and propose treating it by a "nested model." We should match the behavioral design and the statistical model. The design they are analyzing crossed the A and B factors;

	B1	B2	B3
A1			
A2			

so the six cells are A1B2 cells, etc. as expected. Yet L&M propose analysis by a nested model in which B is nested under A. The nested experimental design would look like this:

A1	B1	B2	B3
A2	B4	B5	B6

Now a given level of B occurs within only one level of factor A, so that B is indeed nested within A. This is the legitimate and conventional meaning when you say that B is nested in A. Taking B as a fixed factor, say level of difficulty of the reading material, and A1 and A2 as two methods of teaching reading, such an experiment would make little sense. If B1 is the easiest material, and the difficulty increases up to B6, the experiment is completely confounded, and no interpretations of the A1 versus A2 effect are possible. For this reason most sensible experiments where B is nested in A will be restricted to cases where B is a random factor. Adding an R to indicate a random as opposed to a fixed factor, we may represent the nested design as an RS_n in [RB_3 in (A_2)]. This would clearly indicate that subjects are nested in the six cells, while three different levels of the random factor B are nested in A. B would usually be random classes, schools, or hospitals, etc. Such designs are often referred to as hierarchial designs. This is the conventional use of the word nesting, and of nested experimental designs for which a nested model is appropriate. If L&M wish to propose a misguided "simple effects model" as I previously (1973) labeled it, let us at least not confuse our students by this complete misuse of the word nesting. A major skill in data analysis is matching the model to the experimental design. The fact that L&M adopt a model that others correctly use for nested factors when L&M have crossed factors should give them a cue that they are on weak ground. But to cut the ground out from everybody else by complete misuse of this vital term is intolerable.

So to the Type IV errors. [redacted] If any of you take such errors seriously, take the time to go talk with somebody who does a lot of data analysis in the physical sciences, or a mathematical statistician who does a lot of consulting and data analysis. Tell him how you conducted your last four factor AOV, and ask how he would have analyzed it. Often they will ask questions that are surprising for people trained only in the Kirk, Winer, and Lindquist (1953) tradition. They may ask "Did you really expect to get a four factor interaction and all four three-factor interactions?" If your answer is no, a likely response is "Then why did you include them all in your model?" In the physical sciences the tradition is the opposite of ours. They typically start with the simpliest model that they expect to be consistent with reality. They then fit this model to the data, and start an extensive search of plots of the residuals to detect ways in which their model can be improved. Read Draper and Smith (1966) and note their entire chapter on the examination of residuals. It is not uncommon to generate 10-15 plots of residuals with various predictors. Then read Daniel and Wood (1971) and see additional procedures based on residuals that were invented in the intervening five years. Additional tools are still being invented and are one of the frontiers of applied statistics. All of this work is done so that the model will be changed and improved in the next step of the analysis. If you tell these men that it is a statistical sin to change the model, they will

kick you out of their office as a hopeless clod who is beyond redemption. L&M describe changing models as Type IV errors, and then state "...Type IV errors should be considered in the same vein as Type I and Type II errors when designing an experiment" (1976, p. 64). Clearly L&M consider Type IV errors a form of statistical sin.

I consider the major statistical sin is ignoring power considerations, and in particular throwing away power by use of such methods as the Scheffe' test on an apriori contrast. If L&M are interested in statistic sin, perhaps they should spend some time in church, i.e., go to meetings of full time statisticians. They may be astonished that the church is so full of sinners. They may be even more astonished to find the preacher is teaching the listeners how to sin better, i.e., how to recognize cues indicating a particular model change is desirable. They will also find some of those preachers and listeners are past presidents of the American Statistical Association or similar organizations. Take the L&M articles along when you visit your local statisticians and ask their opinion on the Type IV error sections - not the rest.

Our AOV tradition as illustrated in Lindquist, Kirk, and Winer, is what I call the compulsively complete model method. We are traditionally what Freudians would label as anal retentive. We hate to drop a possible term until we have tested it and found it nonsignificant. If that isn't anal retentive, what is? Thus we start with a four factor model and test the four factor interaction. If our prayers to Pearson and Fisher are answered, this term is not significant, and we proceed to test the three factor interaction terms. When we drop extra terms, we usually do not bother recomputing the error term by pooling the SS_{ABCD} with the prior SS_E. In most good experiments the df to be gained by such pooling is trivial so power consequences are negligible. There are reasonable arguments that can be made against such pooling that suggest

we stick to a pure measure of MS(error) throughout.

I am reasonably sympathetic with this "anal retentive" tradition in the behavioral sciences data analysis and teach it in my early courses. The "build to complex models from a simple start" approach requires greater mathematical sophistication, greater experience, and even some "artistic" skill. Since the median number of publications of APA members still is 1.0, and probably is not very different for AERA members, I prefer to start with techniques that are more readily mastered. However, unlike L&M I am not about to label most of the physical science type data analysts as "sinners" because of their methods. Many of them are excellent data analysts, and I have been occasionally amazed at how much they can obtain from their data with these techniques. However, much success may be due to much smaller error terms than we have.

There is even almost a tradition in linear model type data analysis of reanalyzing somebody else's data with improved models. Draper and Smith (1966) must have at least five or so cases of this. Daniel and Wood (1971) proceed to apply a third model, and another analysis to some of this same data. Do L&M really consider that these eminent statisticians are sinners whose work should be discarded because they repeatedly make Type IV errors? Even in the behavioral sciences we have many articles illustrating the necessity of changing models as you proceed thru an analysis. Appelbaum and Cramer's (1974) discussions of analysis of nonorthogonal AOV designs and Cramer's discussion of multiple regression (1973) are good examples.

In short, L&M's insistence that you should take one model, to love and to cherish, through thick and through thin, through sickness and health, until death do you part is extending the fidelity appropriate to

a marriage to the completely inappropriate situation of a statistical analysis. Their plea to avoid Type IV errors is no more than a plea for absolute rigidity on the part of the statistician. They not only want us to be anal retentive, they absolutely prohibit dropping of terms, therefore their followers must be labeled as anal constipative.

In conclusion, let me share a letter received from Rupert Miller, a professor in the Department of Statistics at Stanford (read quote).

This summarizes the business of a data analyst nicely. Allow me to paraphrase Miller. The aim in data analysis is to extract as much information from the data as possible. Rigid adherence to one technique or one model is foolish. You want to learn whether your model is reasonably correct, and as simple as possible, and only then do we estimate the parameters in the model as accurately as we can.

Thus the data analysis should be flexible, free to change his model just as a good scientist changes his theory to fit the data. I would add that we want to be sure that our model is a reasonable one before we start worrying about such subtitles of statistics as minimum variance estimators, maximum likelihood, etc. L&M's Type IV errors are nothing more than a plea for rigidity in the statistician. In my estimation, we already have a surplus of rigidity, what we need is more flexibility instead.

(Tell "take the bus" analogy and "powder out the window" joke if time remains.)

References

Appelbaum, M. I., & Cramer, E. M. Some problems in the nonorthogonal analysis of variance. Psychological Bulletin, 1974, 81, 335-343.

Bennett, C. A., & Franklin, N. L. Statistical analysis in chemistry and the chemical industry. New York: Wiley, 1954.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cramer, E. M. Significance tests and tests of models in multiple regression. The American Statistician, 1972, 26, 26-30.

Daniel & Wood. Fitting equations to data. New York: Wiley, 1971.

Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Games, P. A. Multiple comparisons of means. American Educational Research Journal, 1971, 8, 531-565.

Games, P. A. Type IV errors revisited. Psychological Bulletin, 1973, 80, 304-307.

Kimball, A. W. Errors of the third kind in statistical consulting. Journal of the American Statistical Association, 1957, 52, 133-142.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Cal.: Brooks/Cole, 1968.

Lee, W. Experimental design and analysis. San Francisco: Freeman & Co., 1975.

Levin, J. R., & Marascuilo, L. A. Type IV errors and interactions. Psychological Bulletin, 1972, 78, 368-374.

Levin, J. R., & Marascuilo, L. A. Type IV errors and Games. Psychological Bulletin, 1973, 80, 308-309.

Li, C. R. Statistical inference. Vol. 1. Ann Arbor: Edwards Bros., 1964.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

Marascuilo, L. A., & Levin, J. R. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. American Educational Research Journal, 1970, 7, 397-421.

Marascuilo, L. A., & Levin, J. R. The simultaneous investigation of interaction and nested hypotheses in two-factor analysis of variance designs. American Educational Research Journal, 1976, 13, 61-65.

Myers, J. L. Fundamentals of experimental design (2nd ed.). Boston: Allyn & Bacon, 1972.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.