DOCUMENT RESUME

ED 093 926 TM 003 744

AUTHOR Lai, Morris K.
TITLE The Case Against Tests of Statistical Significance. Teacher Education Division Publication Series. Report A73-20.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
PUB DATE [73]
NOTE 10p.; Paper presented at the Annual Meeting of the California Educational Research Association (Los Angeles, California, 1973)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Analysis of Variance; *Hypothesis Testing; *Problems; Statistical Analysis; *Tests of Significance

ABSTRACT

The purposes of this paper are to: (1) describe some of the serious shortcomings in the current use of tests of statistical significance, (2) discuss how misuses are perpetuated in some widely used references, and (3) present an alternative significance testing model that overcomes some, but not all, of the shortcomings of the currently used method. For the purposes of this paper, the discussion is restricted to fixed effects analysis of variance (ANOVA) (including t-tests), which is perhaps the most pervasive of the data analyses used by educational researchers. (Author)

TEACHER EDUCATION DIVISION PUBLICATION SERIES

THE CASE AGAINST TESTS OF STATISTICAL SIGNIFICANCE

Morris K. Lai

Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, 1973

REPORT A73-20

FAR WEST LABORATORY FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
1855 Folsom Street, San Francisco, California, 94103, (415) 565-3000

The Case Against Tests of Statistical Significance

Morris K. Lai
Far West Laboratory for Educational Research and Development
1855 Folsom Street, San Francisco, California 94103

Paper presented at the Annual Meeting of the
California Educational Research Association, Los Angeles, 1973.

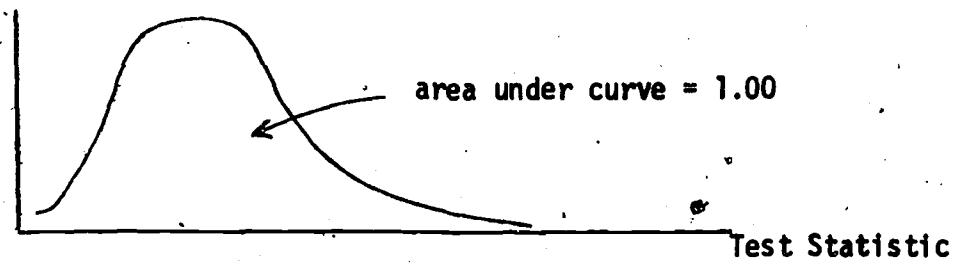
The Case Against Tests of Statistical Significance

Morris Lai
Far West Laboratory for Educational Research and Development

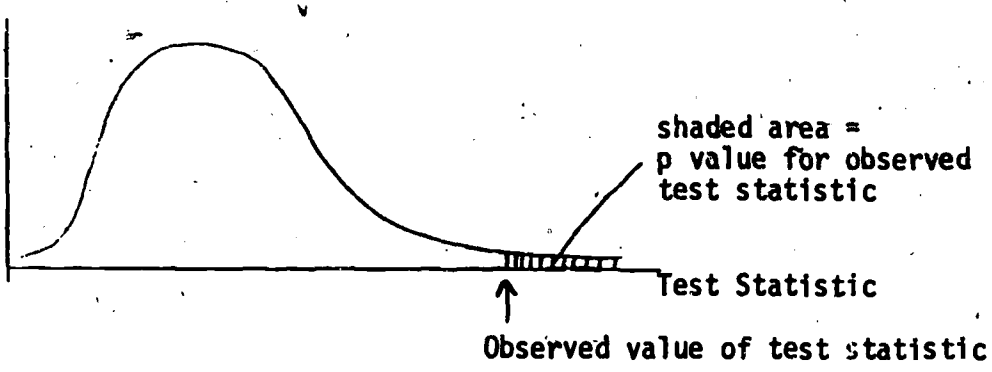
The purpose of this paper is to (1) describe some of the serious shortcomings in the current use of tests of statistical significance, (2) discuss how misuses are perpetuated in some widely used references, and (3) present an alternative significance testing model that overcomes some, but not all, of the shortcomings of the currently used method.

Defining "testing statistical significance"

For the purposes of this paper, the discussion will be restricted to fixed effects analysis of variance (ANOVA) (including t-tests), which is perhaps the most pervasive of the data analyses used by educational researchers. A test of statistical significance is basically a process whereby two or more groups are compared, and for whatever difference is found, a "p value" is calculated which is the probability that a difference that large or larger would have arisen in a sample had the groups been truly equivalent as populations.



Distribution of test statistic when groups are equivalent in the population



For observed test statistics that are sufficiently large, the p values are correspondingly small (i.e., statistically significant).

Random assignment

Such a model requires, to start off with, random sampling. If assignment to treatment is not random, then a test of significance is inappropriate (Morrison & Henkel, 1969).

Type I error rate

Nearly every textbook on inferential statistics discusses the concept of Type I and Type II errors. Despite warnings from Horst (1966), Skipper et al. (1967), and Winer (1971) about the inappropriateness of endowing Type I error rates of .05 and .01 with some sort of sacredness, the prevalence of such sacredness is well known (e.g., the APA Publication manual advocates one asterisk for $p < .05$ and two asterisks for $p < .01$).

Practical or educational significance

It is popular today to exhibit some enlightenment by emphasizing that statistical significance does not necessarily imply practical or educational significance. Yet in Guilford's (1956) widely used textbook we find the following quote: (p. 275)

> The F ratio for machines is significant beyond the .01 level, leaving us with considerable confidence that the machine differences, as such, have a real bearing upon the difficulty of the task.

Such a significant F could have resulted where the differences were trivial in the practical sense. Another misuse of p levels occurs when researchers use significance levels to compare results from several studies (e.g., Eysenck, 1960; Bracht, 1970).

Type II error rates, power, and accepting null hypotheses

Type II error rates and power calculations are less familiar to researchers. Anyone who accepts a null hypothesis, without knowing the power of the statistical test, is liable to have a huge Type II error rate. Yet Popham (1967) in his text writes "...hypothesis under consideration is either accepted or rejected." Glass and Stanley (1970 also mislead their readers by advocating, without consideration of power, the acceptance of the null hypothesis when statistical significance is not attained. Other writers who advocate (inappropriately) the accepting of null hypotheses if a significant statistic is not observed include Walter and Lev (1953) Guilford (1956), and Kirk (1968).

It is possible to prove algebraically that for a predetermined level of significance, there exist normal distributions such that the F or t statistic will not be significant, but the size of the effects will be larger than any predetermined number. As such, a researcher who accepts a null hypothesis without knowing the power of the test may be calling a very large difference a "zero difference." McNemar's (1962) suggestion of using three regions (acceptance, suspended judgment, and rejection), depending on the p level, does not overcome this objection.

Sample size

Another problem that I will discuss is determining sample size. Any scientist appreciates the fact that the larger the sample, the more information one has. Aside from cost-benefit considerations and manageability, it is illogical to say that a smaller sample is more desirable than a larger one; for example, Hays (1963) clearly states that for precision, the bigger the sample size the better. Yet on the next page (p. 324) he suggests that the researcher ask the following question: "Is the sample size large enough to give confidence that the big associations will indeed show up, while being small enough so that trivial associations will be excluded from significance?" If a procedure is such that it results in worry about whether a sample size is small enough, then surely something is seriously wrong with that procedure.

Appropriate null hypotheses

The last problem I will discuss deals with null hypotheses. The unquestioning acceptance of always using a zero difference null hypothesis has been criticized by several writers (e.g., Grant (1962); Kerlinger (1964); Cohen (1969). Dixon and Massey (1969) and Pena (1970) have both presented a procedure for testing non-zero null hypotheses for the two sample case. The incorporation of a predetermined minimum practical difference into the null hypothesis (now non-zero) ties in the statistical and practical significance. By means of this rarely used procedure, a researcher can state more appropriate null hypotheses. Instead of asking if there is a difference at all, researchers usually should be asking whether or not there is an educational or practical difference. Instead of asking whether

a Datsun gets better mileage than a Cadillac, we should be asking how many more gallons a Datsun gets and whether this difference was of practical importance. Likewise instead of asking whether one group has scored higher than another, we should be asking how much higher one group has scored than another and if this difference is of practical or educational importance.

Summary

In summary, well respected writers have suggested that researchers do the following (1) test null hypotheses that are usually inappropriate, (2) accept these null hypotheses without regard to power (and possibly have huge Type II errors), (3) use arbitrary (sacred) rejection probability levels of .05 and .01, and (4) be careful in not getting too large a sample size.

These misleading (inappropriate) recommendations are interrelated in that their disappearance would be highly correlated with the elimination of tests of significance. But change comes slowly and I propose an analysis of variance methodology that gets rid of (1) and (4) (inappropriate null hypotheses and the illogical concept of a sample being too large.)

Noncentral analysis of variance

The method can perhaps be best understood in terms of its being an extension of the two sample case which has been described by Dixon and Massey (1969). The analog to the minimum practical difference, is δ, the noncentrality parameter of the noncentral F distribution. Just as the ordinary F distribution is associated with a zero difference null hypothesis, the noncentral F distribution is associated with a non-zero null hypothesis. Minimum practical differences are now stated in terms of average differences between groups.

The derivation of the noncentral ANOVA model is complex and will be presented in more detail in another paper. The use, however, is rather simple. Having determined the minimum practical difference, a researcher need only use a table to determine the noncentrality parameters. He then rejects the (nonzero) null hypothesis if his observed F statistic exceeds $F_{\nu_1, \nu_2, \delta}(1-\alpha)$, where ν_1 and ν_2 are the usual parameters that determine the central F distribution, δ is the noncentrality parameter and α is the Type I error rate chosen.

Such a procedure results in an appropriate adjustment for sample size. Thus, statistical significance is not attainable by merely increasing the sample size. The illogical concept of too large a sample no longer exists. At the same time, appropriate null hypotheses are being tested.

Bibliography

American Psychological Association, Publication manual. Washington, D. C.: APA, 1967.

Bracht, G. "Experimental factors related to aptitude-treatment interactions." Review of Educational Research, 1970, 40(5), 627-645.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Dixon, W., and Massey, F. Jr. Introduction to statistical analysis. New York: McGraw-Hill, 1969.

Eysenck, H. J. "The concept of statistical significance and the controversy about one-tailed tests." Psychological Review, 1960, 67(4), 269-271.

Glass, G., and Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Grant, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, 1962, 69, 54-61.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1956.

Hays, W. L. Statistics. New York: Holt, Rinehart & Winston, 1963.

Horst, P. Psychological measurement and prediction. Belmont, California: Wadsworth, 1966.

Kerlinger, F. Foundations of behavioral research. New York: Holt, Rinehart & Winston, 1964.

Kirk, R. Experimental design: procedures for the behavioral sciences. Belmont, California: Brooks/Cole, 1968.

McNemar, Q. Psychological statistics. New York: John Wiley & Sons, 1962.

Morrison, D. E., and Henkel, R. E. Significance tests reconsidered. The American Sociologist, 1969, 4, 131-140.

Morrison, D. E., and Henkel, R. E. The significance test controversy. Chicago: Aldine, 1970.

Pena, D. A significant difference of opinion with the Coats position. Educational Researcher, 1970, 21, 9-10.

Popham, W. J. Educational statistics. New York: Harper & Row, 1967.

Rosenkrantz, R. The significance test controversy. Educational Researcher, 1972, 1(12), 10-14.

Skipper, J. K., Guenther, A. C., and Nass, G. The sacredness of .05: a note concerning the uses of statistical levels of significance in social science. The American Sociologist, 1967, 1, 16-18.

Walker, H. M., & Lev, J. Statistical inference. New York: Henry Holt & Company, 1953.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.