DOCUMENT RESUME

EL 303 514 TM 012 775

AUTHOR Keaster, Richard D.
TITLE Statistical Significance Testing: From Routine to Ritual.
PUB DATE Nov 88
NOTE 15p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Louisville, KY, November 9-11, 1988).
PUB TYPE Speeches/Conference Papers (150) - Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Assessment; Research Design; Research Methodology; *Research Problems; Statistical Analysis; *Statistical Significance; Statistics

ABSTRACT

An explanation of the misuse of statistical significance testing and the true meaning of "significance" is offered. Literature about the criticism of current practices of researchers and publications is reviewed in the context of tests of significance. The problem under consideration occurs when researchers attempt to do more than just establish that a relationship has been observed. More often than not, too many researchers assume that the difference, and even the size of the difference, proves or at least confirms the research hypothesis. Statistical significance is not a measure of "substantive' significance or what might be called scientific importance. Significance testing was designed to yield yes/no decisions. It is suggested that authors or research projects should not try to interpret the magnitudes of their significance findings. Significance testing must be returned to its proper place in the scientific process. (SLD)

STATISTICAL SIGNIFICANCE TESTING:

FROM ROUTINE TO RITUAL

Richard D. Keaster

University of New Orleans

Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY, November 9, 1988.

ABSTRACT

The literature regarding the criticism of current researcher and publication practices is reviewed in the context of tests of significance. An explanation of the misuse of statistical significance testing and the true meaning of "significance" is offered. Contentions are supported by statements of argument and suggestions are proposed to govern future practices regarding the interpretation of research results and their publication.

A "ritual," as defined by Webster's Third New International Dictionary, is "any practice done or regularly repeated in a set precise manner so as to satisfy one's sense of fitness and often felt to have a symbolic or quasi-symbolic significance." The present paper attempts to demonstrate how attitudes toward the null hypothesis and misuses of results of statistical significance have been given "ritual" status and, as a result, call into question the integrity of a good deal of "scientific research."

Ever since Sir Ronald Fischer promulgated the process of null hypothesis testing, a continuing controversy has brewed over statistical significance and its proper place within the scientific method. Carver (1978, p. 389-390) suggests that attitudes toward the importance of significance testing imply that a "corrupted" model of the scientific method has been adopted by the research community. Under the "true" method (in a differences comparison),

> . . . (A) research hypothesis is tested by collecting data and then comparing the results with those predicted from the research hypothesis. . . . (I)f the difference between the mean of the experimental group and the mean of the control group is in accordance with what was predicted by the research hypothesis, then this constitutes evidence in favor of the research hypothesis.

Carver (1978) goes so far as to say that "Educational research would be better off if it stopped testing its results for statistical significance." This opinion is supported by a

number of other individuals. Cronbach (1975 p. 124) asserted that "the time has arrived to exorcise the null hypothesis" and Shulman (1970, p. 389) concurred demanding that "educational researchers . . . divest themselves of the yoke of statistical hypothesis testing." Bakan (1966, p. 436) referred to this practice, as demonstrated in psychological research, as a kind of "essential mindlessness." More recent discussion of these issues is provided by Huberty (1987) and Kupfersmid (1988).

Despite such overt criticism of what can only be considered as "research malpractice," the research community as a whole continues to treat "statistical" significance as if it were the same as "substantive" significance. A brief explanation of what statistical significance actually is and what it can and cannot do for the inquiring researcher is in order.

In an experiment where a comparison of means is conducted, two samples are drawn from a population providing for randomness, stratification, and all other devices for controlling for bias. One group (the experimental) receives a treatment and the other group (the control) does not. Following the treatment, the mean scores on some test of the two groups (depending on the design and what exactly is being tested) are compared for differences.

The null hypothesis assumes that the two means will not be different to any substantial degree, thus implying the two groups represent the same population and that no effect has been witnessed. As Carver (1978, p. 381) states,

> The null hypothesis states that the experimental group and the control group are not different . . . and that

any difference found between their means is due to sampling fluctuation. Statistical significance testing sets up a straw man, the null hypothesis, and tries to knock him down. We hypothesize that the two means represent the same population and that sampling or chance alone can explain any difference we find between the two means.

Taking these things into consideration, the researcher can mathematically deduce exactly how often differences as large as or larger than what was found would occur by the chance of sampling error. If larger differences are found, the null hypothesis (which states that there will be essentially no difference) would be rejected, implying that the treatment did have an effect or that there was some relationship between the treatment and the differences observed in the means. Here is where "statistical significance" ends.

The problem under consideration occurs when the researcher attempts to do more than just establish that a relationship has been observed. More often than not, too many researchers assume that the difference, and even the size of the difference, proves, or at least confirms the research hypothesis.

Gold (1969, p. 43) suggests that:

A test of significance under the best of circumstances provides only an index of reliability, restricted by time, place, and people, so that we are in fact dealing with unique historical knowledge. Statistical analysis can be considered only a preliminary screening that any hypothesis must pass to merit further investigation.

Significance testing was designed to yield "yes-no" decisions. Consequently, authors of research projects should refrain from the temptation of interpreting the magnitudes of their significance findings (Lykken, 1968; Thompson, 1988).

Schneider and Darcy (1984) list seven factors that determine the outcome of significance tests:

1) Actual strength of impact
2) Number of cases used in the study
3) Variation among cases on relevant variables
4) The complexity of the analysis (degrees of freedom)
5) The appropriateness of the statistical measures and tests used
6) The hypothesis tested
7) The significance level chosen

Only one of these seven deals with the impact of the outcome, and that impact cannot be measured simply by looking at a test designed for a "yes-no" decision.

"Statistical" significance is not a measure of "substantive" significance or what might be referred to as "scientific importance." To assign greater meaning than what is warranted commits one to a line of thought where assertions and theories attain "factual" status. Kish (1959, p. 336) confirms these assertions by stating, "*Significance* should stand for meaning and refer to substantive matter. . ." and suggests dropping the phrase "test of significance." This error is condemned in the scientific method, yet much of our practice, and worse yet, much of what is published in research literature, supports the fact

that these "leaps" in logic are apparently not only condoned but actually encouraged.

One view suggests that significance testing has been "saddled" with too much responsibility. Bakan (1966, p. 423) notes:

> The argument is . . . that the test of significance has been carrying too much of the burden of scientific inference. Wise and ingenious investigators can find their way to reasonable conclusions from data because and in spite of their procedures. Too often, however, even wise and ingenious investigators, for varieties of reasons not the least of which are the editorial policies of our major psychological journals. . . tend to credit the test of significance with properties it does not have.

Thompson (1988) suggests four current practices which fail to demonstrate the acknowledgment of the limitations of significance testing. First, he asserts that counseling and psychological literature demonstrates an apparent bias against articles that do not report (statistically) significant results. This is supported by Greenwald (1975) and Atkinson, Furlong, and Wampole (1982). Second, citing Cohen (1979), he asserts that readers of such literature perceive articles reporting significant results more favorably than those that fail to do so. Third, editors of counseling and psychological journals possess this same attitude toward articles that do and do not report significant results, as asserted by Carver (1975) and confirmed by Greenwald (1975) and Atkinson, Furlong, and Wampole (1982).

Finally, as a result of these attitudes and practices, authors refrain from submitting articles that cannot report statistical significance, or they hesitate even pursuing lines of inquiry based on these results (Greenwald, 1975).

Atkinson, Furlong, and Wampole (1982) note the considerable influence that psychological journals exert over training, practice, and research. Time spent in the reading of these journals for preparation in future research by graduate and undergraduate students, as well as the journals' usage by practicing professionals, stresses the importance of maintaining standards of style and content.

But research conducted by Atkinson, Furlong, and Wampole (1982) was designed to determine if the level of statistical significance alone affected the evaluation and recommendation of manuscripts for publishing. Their conclusions support earlier reports (Bakan, 1966; Craig, Eison, & Metze, 1976; Greenwald, 1975; Lykken, 1968; Selvin, 1957) that these journal editors, who function as the "ultimate 'teachers,'" have encouraged practices which are "patently wrong" (Bakan, 1966, p. 430) through their publishing policies and practices.

An interesting, yet similarly unfortunate, "syndrome of indelibility" results from these editorial practices. Type I errors (rejecting the null hypothesis when it is actually true) can theoretically occur in one of every twenty studies (using the logic and customary levels of significance testing). When Type I errors are made, significance is found and apparently encourages publication of the study. Bakan (1966, p. 427) asserts, "The

damage to the scientific enterprise is compounded by the fact that the publication of 'significant' results tends to stop further investigation." Bakan argues that if the practice of publishing reports of "significance" fostered attempts at replication and further investigation, the damage of the practice would be lessened. However, the opposite is the case.

Bakan (1966, p. 427-428) also argues that "highly significant" studies have the tendency of appearing definitive and become "archived:"

> Even the strict repetition of an experiment and not getting significance in the same way does not speak against the result already reported in the literature. For failing to get significance, speaking strictly within the inference model, only means that that experiment is inconclusive; whereas the study already reported in the literature, with a low *p* value, is regarded as conclusive.

As a result, we, as researchers, tend to behave in a manner that fails to be self-monitoring, let alone self-correcting.

Suggestions for Improvement

The foregoing argument implies a very serious problem that not only currently exists in our research and publishing practices, but has existed for a number of years. It is time to heed advice for correcting the problems previously noted.

Of Schneider and Darcy's (1984) seven features that affect the outcome of a significance test, Thompson (1988) notes that the one having the greatest impact is the size of the sample.

The suggestion is made that when authors interpret significance, they should consider sample size. Where significance was found, authors should identify the smallest sample size at which the result would have remained statistically significant; where significance was not found, authors should identify the sample size at which the results would have achieved that status.

Another suggestion would be to examine the "effects size." The basic question to be answered when conducting research is to establish "how much of the dependent variable is accounted for by the independent variable," or " what proportion of the variance in the dependent variable is explained by the observed effect?" As Thompson (1988, p. 147) notes, ". . . using effects sizes in interpretation focuses interpretation on what the researcher really cares about (e.g., the issue of noteworthiness of results)."

Lykken (1968, p. 158-159) summarizes his contentions by saying,

> The value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on. Ideally (though not realistically), all experiments would be replicated before publication. . .

These statements indicate there is much more to significance than just statistical testing.

Selvin (1957, p. 527) concludes his arguments on the use of significance testing in sociology by asserting:

> Sociologists would do better to re-examine their purposes in using the tests and to try to devise better methods of achieving these purposes than to continue to resort to techniques that are at best misleading for the kinds of empirical research in which they are principally engaged.

Carver (1978), who advocates as a last resort the abolition of significance testing, states,

> If one could no longer use statistical significance to determine the "significance" of a difference, researchers would be forced to use designs that more clearly reveal the scientific importance of a difference. Without statistical significance, researchers will be forced to grapple with the problems of scientific inference instead of those associated with statistical testing.

Bakan (1966, p. 436) quotes Karl Pearson as saying that, "higher statistics (are) only common sense reduced to numerical appreciation. However, that base in common sense must be maintained with vigilance."

Conclusion

The obvious conclusion that must be drawn would be that significance testing should be returned to its proper place of importance in the scientific process. As Bolles (1962, p. 645) contends,

> Our present day over-reliance upon statistical hypothesis testing is apt to obscure this feature (probability conclusions) of the scientific enterprise. We have almost come to believe that an assertion about the nature of t! empirical world can be validated. . . in one stroke if the data demonstrate statistical significance. Is it any wonder then that our use of statistical hypothesis testing is rapidly passing from routine to ritual?

The findings and conclusions of this paper indicate that the "passing" is complete. In the 25 years since the printing of Bolles' statement, our publishing practices of over-emphasizing the importance of statistical significance has truly moved the practice from the realm of "routine" to that of "ritual."

REFERENCES

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process. Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29 (2), 189-194.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66 (6), 423-437.

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. Psychological Reports, 11, 639-645.

Carver, R. P. (1978). The case against significance testing. Harvard Educational Review, 48 (3), 378-399.

Cohen, L. H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. Journal of Consulting and Clinical Psychology, 47, 421-423.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and ω^2. Bulletin of the Psychonomic Society, 7 (3), 280-282.

Cronbach, L. J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Gold, D. (1969). Statistical tests and substantive significance. The American Sociologist, 4, 42-46.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82 (1), 1-20.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16 (8), 4-9.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Kish, L. (1959). Some statistical problems in research design. American Sociological Review, 24, 328-338.

Lykken, D. T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70 (3), 151-159.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8 (4), 573-583.

Selvin, H. C. (1957). A critique of tests of significance in survey research. American Sociological Review, 22, 519-527.

Shulman, L. S. (1970). Reconstruction of educational research. Review of Educational Research, 40, 371-393.

Thompson, B. (1988). A note about significance testing. Measurement and Evaluation in Counseling and Development, 20 (4), 146-148.

REFERENCES

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29 (2), 189-194.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66 (6), 423-437.

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. Psychological Reports, 11, 639-645.

Carver, R. P. (1978). The case against significance testing. Harvard Educational Review, 48 (3), 378-399.

Cohen, L. H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. Journal of Consulting and Clinical Psychology, 47, 421-423.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and ω^2. Bulletin of the Psychonomic Society, 7 (3), 280-282.

Cronbach, L. J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Gold, D. (1969). Statistical tests and substantive significance. The American Sociologist, 4, 42-46.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82 (1), 1-20.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16 (8), 4-9.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Kish, L. (1959). Some statistical problems in research design. American Sociological Review, 24, 328-338.

Lykken, D. T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70 (3), 151-159.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8 (4), 573-583.

Selvin, H. C. (1957). A critique of tests of significance in survey research. American Sociological Review, 22, 519-527.

Shulman, L. S. (1970). Reconstruction of educational research. Review of Educational Research, 40, 371-393.

Thompson, B. (1988). A note about significance testing. Measurement and Evaluation in Counseling and Development, 20 (4), 146-148.