INADEQUACIES OF SIGNIFICANCE TESTS IN EDUCATIONAL RESEARCH

M. S. Lalithamma
Masoomeh Khosravi

Tests of statistical significance are a common tool of quantitative research. The goal of these tests is to determine the probability that the results observed could have been the result of chance where the null hypothesis is exactly true in the population. The tests, however, have been used widely in the past 70 years as "proof" of the effectiveness of a given treatment or the existence of noteworthy relationship between variables. Researchers frequently have used the results of a significance test to assess whether their findings are valid. Although the value of significance testing continues to be debated, most experts believe it is inappropriate to make a judgment about the truth of a relationship between variables based upon the results of a significance test. In one sense, there is nothing wrong with significance testing. Rather, the inadequacies of the interpretation and application of the results are at issue. The purpose of this article is to summarize the criticisms and limitations that have been leveled at significance testing and also suggest effect size as supplement statistical significant procedures.

INTRODUCTION

Scholars have used statistical testing for research purposes since the early 1700s (Huberty 1993). In the past 300 years, applications of statistical testing have advanced considerably, most noticeably with the advent of the computer and recent technological advances. However, much of today's statistical testing is based on the same logic used in the first statistical tests and advanced in the early twentieth century through the work of Fisher, Neyman, and the Pearson family. Specifically, significance testing and hypothesis testing have remained at the cornerstone of research papers and the teaching of introductory statistics courses.

Currently, we are in an era where the value of statistical significance testing is being challenged by many researchers. Research methodology literature in recent years has included a full frontal assault on statistical significance testing. The assault is based on whether or not statistical significance testing has value in answering a research question posed by the investigators. One of the most important contemporary criticisms emphasises the need that researchers must evaluate the practical importance of results, along with testing for statistical significance. Kirk (1996) agreed that statistical significance testing is a necessary part of a statistical analysis. However, he asserted that the time had come to include practical significance in the results. He recommended the use of statistical significance testing; however, it must be considered in combination with other criteria. Specifically, statistical significance is one of the three criteria that must be demonstrated to establish a position empirically; the other two are being practical significance and replicability. This paper considers both use of and problem with statistical significance testing. First, some relevant issues related to statistical significance testing in research are briefly reviewed and then the major criticisms of statistical significance tests are explained. Finally, in response to the criticisms, reporting effect size which should be done in conjunction with statistical significance testing has been suggested.

Use of statistical significance testing in research

It is important to have a good understanding about what basic purpose statistical significance testing provides for researchers. The fundamental concept underlying statistical significance testing is sampling variation: from a population with known parameters (e.g., known population mean), sample statistics (e.g., observed sample mean) will vary around the population parameter to certain extent. How much sampling variation can there be? How likely will an observed sample statistic (e.g., sample mean of 68) can occur due to sampling variability (i.e., by chance) for a given population parameter (e.g., population mean of 80)? In a nutshell, statistical significance testing is conducted to evaluate the viability of null hypothesis by assessing how likely some observed sample statistic could have occurred as the result of random sampling variation for a given population parameter. More specifically, statistical significance testing answers the question: what is the probability of obtaining an observed sample statistic for a given or known population parameter?

Assuming that there exist two treatment conditions, A and B (A represents a new instructional approach in teaching mathematics, while B, represents the conventional instructional approach currently in use). The program evaluation team is interested in knowing if A is better and more effective than B in teaching maths to children. The null hypothesis in this situation is that A and B are equal, i.e., students under A and B will learn equally well. Obviously, because of sampling variation, the two samples (one under A, and the other under B) typically will not have the same statistics, even if A is needed the same as B. The question becomes: how different the sample statistics should be between A and B samples when one can say with confidence that A is different from B in effectiveness. Given the null hypothesis of no difference between A and B treatments, smaller observed difference between A and B samples is more likely to occur than larger observed difference between two. When the difference between the two samples become sufficiently large relative to the theoretical random sampling variation such that it becomes highly unlikely if A and B are equally effective (null hypothesis of no difference), one concludes that the observed result is very unlikely to have occurred if the null hypothesis is indeed true. As a result, the null hypothesis of no difference is rejected and it is concluded that A and B are not the same in their effectiveness in teaching maths.

It may be noted that in statistical significance testing, all that is assessed is the probability of obtaining the sample data (D) if the null hypothesis (H0) is true. If p is sufficiently small (e.g., smaller than .05 or .01), the null hypothesis will be considered not viable, and will be rejected. The rejection of the null hypothesis reveals that the random sampling variability is the unlikely explanation for the observed statistical results, but it gives no indication about importance of the obtained statistical results. Going back to the example of A and B approaches in teaching mathematics, rejection of the null hypothesis simply reveals that it is unlikely that A and B are equally effective, but it does not give one any indication about how much more effective A is than B, or vice versa . The real meaning of statistical significance testing, however, has often been lost in research practice, and the importance of statistical significance tends to be greatly exaggerated.

COMMON CRITICISMS ABOUT SIGNIFICANCE TESTING

The most frequently expressed concern is simply that statistical significance testing identifies statistically significant results but not necessarily practically significant results. This criticism is partly a product of the relationship between the size of a study (number of participants or degree of freedom) and statistical significance. Even the smallest relationships can become statistically significant, if a large enough sample is used for the study. Thus, the argument goes, many studies are getting published and taking on disproportionate importance because they demonstrate statistical significance, even though the magnitude of the effect measured has no practical value.

Biskin (1998) pointed out that increased sample size will eventually yield statistically significance only if the null hypothesis is false. Further more because the null hypothesis refers to population parameters. They are truly experimental studies in which the null hypothesis is true. However, Vacha-Haase and Thompson (1998) have questioned whether this inference can be extended from a theoretical population to actual sample values. They and others have argued that in practice the null hypothesis is essentially always false. Thus, statistical significance testing becomes a tautological exercise in demonstrating evidence for what is already known.

The second criticism of the use of statistical significance testing is that the null hypothesis, which is fundamental to all statistical significance testing, is often misunderstood and misinterpreted (Kirk 1996). The typical null hypothesis assumes in advance that there are no differences between groups or, in the case of continuous variables, there is no relationship between the variables. The significance test then determines the probability that the reported data would occur given that there is no relationship. However, generally investigators do not want to know this probability. Instead, it would be much more useful to know the probability that there is no relationship, given the reported data. The probability that a researcher's null hypothesis is false, given some set of data, may be quite different from the probability that these data would occur, given the null hypothesis. Ottenbacher (1989) pointed out that this error results from a failure to consider Type II errors. Type II errors result from failing to reject a null hypothesis even though it is true. The probability to Type I errors, which is controlled in conventional significance testing, does not imply that the probability of Type II errors is controlled at similar levels. Cohen (1994) noted in this regard that significance testing "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does".

A third criticism has to do with the a priori selection of an alpha (α) level against which the probability level for each test statistic is to be compared .Whereas conventional choices of alpha levels, such as 0.5, are commonly used in an effort to balance the application of a study's power toward avoiding both Type I and Type II errors (Olejnik 1984; Ottenbacher 1989). Most authors make little effort to actually assess the power as irrelevant (Rosenthal 1979). Moreover, the selection of a specific alpha level imposes an artificial dichotomy on a static (p) that is continuous (Kirk 1996; Thompson 1997; Young 1993). The practical difference between calculated probability of .049 as opposed to one of .051 is certainly not as dramatic as the dichotomous decision that only the former

result is statistically significant, with all that it may imply, whereas the other is not. Frustration with this arbitrary dichotomy may encourage authors to refer to some results as “nearly significant” or “approaching significance.

A fourth criticism of the use of significance testing involves misuse of the results. Perhaps as a consequence of the combination of previously cited criticisms, some authors seem to associate significance testing with replicability or reliability (Schmidt & Hunter 1995; Thompson 1996; 1997; Vacha Haase & Nilsson 1998), which leads to the assumption that a p value of .001, for example, is somehow more important or more impressive than a p value of .05. Certainly, p values are not a measure of the likelihood that a given result will be replicated (Cohen 1994). Nevertheless, getting a very small p value often leads to the potentially misleading description of a result as “highly” significant or as evidence of a “strong effect”, in spite of the fact that p level does not imply the strength of the relationship (Friedman 1968; Vacha-Haase & Thompson 1998).

Schmidt and Hunter (1995) also cautioned against another all too common error when reporting nonsignificant results which is also related to the magnitude of the p value. That is, some authors infer incorrectly that nonsignificance implies that there is no effect. Clearly, a nonsignificant result only indicates that the data being tested do not provide adequate evidence to reject the null hypothesis, given a particular alpha level. The nonsignificant result does not demonstrate that the null hypothesis is true.

Two additional, somewhat more technical criticisms were raised by Thompson (1993). The first criticism involves hierarchical testing within the same data set. Given the recommendation that higher order interactions should be tested in factorial ANOVA studies before main effects (Keppel 1991), Thompson has reminded that each of these tests may represent very different distributions of the samples size across means. These differences could result in very different power to detect differences for each test. Thus, whatever a significant result may mean in an omnibus test that includes the entire sample means, it may mean something very different for a main effect or for some other specific comparison in the same data.

The second technical concern of Thompson (1993) about significance testing relates to the relationship between the sample size and the assumptions on which significance testing is based. For example, ANOVA assumes homogeneity of variances, and ANCOVA additionally assumes homogeneity of regression. In testing these assumptions, investigators conduct significance tests in hopes of not rejecting the null hypothesis. Ironically, the same large sample size that provides power against Type II errors will also increase the likelihood that the null hypothesis rejected, making the use of those significance tests more questionable.

At this point, it should be clear that the objections to the use of significance testing are intimately interrelated. They certainly could have been organized and grouped differently here. Nevertheless, they also represent important concerns for investigators who want their results to be significant in the sense of having practical value.

RECOMMENDED CHANGE IN PRACTICE (SUPPLEMENTING THE STATISTICAL SIGNIFICANT TEST)

Some authors due to criticisms noted previously, have recommended the complete elimination of significance testing (Carver 1993; Morse 1998; Schmidt & Hunter 1995). However, most have taken the more moderate view that significance testing should be supplemented with or placed in the context of additional information. Most regrettably, however, empirical studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "encouraging" effect size reporting has not appreciably affected actual reporting practices (Kirk 1996). Due to this lack of change, authors have voiced stronger opinions concerning the emphasized recommendation (Thompson 1996). 'Effect size' is simply a way of quantifying the size of the difference between two groups. It is easy to calculate, readily understood and can be applied to any measured outcome in Education or Social Science. It is particularly valuable for quantifying the effectiveness of a particular intervention, relative to some comparison. It allows us to move beyond the simplistic, 'Does it work or not?' to the far more sophisticated, 'How well does it work in a range of contexts?' Moreover, by placing the emphasis on the most important aspect of an intervention – the size of the effect – rather than its statistical significance (which conflates effect size and sample size), it promotes a more scientific approach to the accumulation of knowledge. For these reasons, effect size is an important tool in reporting and interpreting effectiveness.

CONCLUSION

Kirk (1996) recently noted that, our science has paid a high price for its ritualistic adherence to null hypothesis significance testing. The overuse and misinterpretation of statistical tests has been frequently decried in the literature. Nevertheless, the use of statistical significance tests remains common, and empirical studies reflect even an increased use of these methods. Use of statistical significance testing as an introduction and foundation for the discussion in this paper reviewed and followed by major criticisms of statistical significance tests. A quick perusal of the criticisms against statistical significance testing cited in present paper, confirm that more appropriate strategies such as effect size reporting which is already discussed and attention to practical significance must be taken to supplement statistical significance tests. Hence, researchers should be encouraged to analyze their results more carefully and demonstrate practically their study outcomes. It is hoped that researchers will be able to find out that the criticisms are having a noticeable impact on educational researches, planning of analysis and reporting of quantitative results.

REFERENCE

Biskin, B. (1998) Comment on significance testing. *Measurement and Evaluation in Counseling and Development* 31, 58–62.

Carver, R. P. (1993) The case against statistical significance testing, revisited. *Journal of Experimental Education* 61, 287–292.

Cohen, J. (1994) The earth is round ($p < .05$). *American Psychologist* 49, 997–1003.

Friedman, H. (1968) Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin* 70, 245– 251.

Huberty, C. J. (1993) Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education* 61,4, 317-333.
Keppel, G. (1991) *Design and analysis: A researcher's handbook.* Prentice Hall., Englewood Cliffs.
Kirk, R. E. (1996) Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 56, 746–759.
Morse, D. T. (1998) Minsize: A computer program for obtaining minimum sample size as an indicator of effect size. *Educational and Psychological Measurement* 58, 142–153.
Olejnik, S. F. (1984) Planning educational research: Determining the necessary sample size. *Journal of Experimental Education* 53, 40–48.
Ottenbacher, K. J. (1989) Statistical conclusion validity of early intervention research with handicapped children. *Exceptional Children* 55, 534–540.
Plucker, J. A. (1997) Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review* 2, 122–126.
Rosenthal, R. (1979)The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86, 638–641.
Schmidt,F.L.,& Hunter, J. E. (1995) The impact of data-analysis methods on cumulative research knowledge: Statistical significance testing, confidence intervals, and meta-analyses. *Evaluation and the Health Professions* 18, 408–427.
Snyder, P. A., & Lawson, S. (1993) Evaluating results using corrected and uncorrected uncorrected effect size estimates. *Journal of Experimental Education* 61, 334–349.
Thompson, B. (1993) The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education* 61, 361–377.
Thompson, B. (1996) AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher* 25, 2, 26–30.
Thompson, B. (1997) Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher* 26, 5, 29–32.
Vacha-Haase, T., & Nilsson, J. E. (1998) Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling and Development* 31, 46–57.
Vacha-Haase, T., & Thompson, B. (1998) Further comments on statistical significance tests. *Measurement and Evaluation in Counseling and Development* 31, 63–67.
Young, M. A. (1993) Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research* 36, 644–656.

****