Mid-Western Educational Researcher

Volume 12 | Issue 2 Article 2

1999

The Use of Tests of Statistical Significance

Thomas R. Knapp
The Ohio State University

The Use of Tests of Statistical Significance

Thomas R. Knapp
Ohio State University

Abstract

This article summarizes the author's views regarding the appropriate use of significance tests, especially in the context of regression analysis, which is the most commonly-encountered statistical technique in education and related disciplines. The article also includes a brief discussion of the use of power analysis after a study has been carried out.

Although statistical significance tests have come under repeated attacks for several years, most recently in psychology by Jacob Cohen (1994), Frank Schmidt (1996), and others, there are times when they should be used and there are times when they should not be used. What follows is an attempt to identify those times as far as educational research is concerned.[1]

A brief history of the controversy, 1970-1998

In 1970 there appeared a book edited by sociologists Denton Morrison and Ramon Henkel, entitled *The Significance Test Controversy*. That book consisted of chapters written by people on both sides of the issue, but most of the authors were "con", i.e., they had little or nothing good to say about significance tests. Several of those chapters had originally appeared elsewhere in books or as journal articles, and some of the comments were downright nasty. In his chapter, for example, Paul Meehl characterized the researcher who uses significance tests as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (Meehl, 1970, p. 265).

For the next couple of decades things were relatively quiet, except for the occasional raising of a few new voices (e.g., Carver, 1978). Significance tests continued to be used by researchers who felt they were warranted and continued to be eschewed by researchers who felt they were not. Then in the 90s, prompted by articles written by Cohen (1990, 1994) and Schmidt (1992, 1996), the controversy was rekindled. It led to the creation of a task force in psychology to deal with the matter and to the publication in 1997 of another entire book devoted to the "pros" and "cons" of significance testing, edited by Lisa Harlow, Stanley Mulaik, and James Steiger, entitled *What If There Were No Significance Tests?* (See Levin, 1998 and Thompson, 1998 for reviews of, and reactions to, that book.) Schmidt had advocated the discontinuation of **all** significance tests in favor of confidence intervals around obtained effect sizes, and the discontinuation of **all** narrative literature reviews in favor of meta-analyses for pooling results across studies. At the time of the writing of this article—Autumn, 1998—the APA Task Force had not issued its final report, but its interim report in 1997 suggested that Schmidt's extreme positions would not be supported.

The situation in educational research has closely paralleled the recent developments in psychology. Starting in 1993 with an entire issue of the *Journal of Experimental Education* devoted to the topic of significance testing (again, "pros" and "cons", but mostly "cons"—see esp. Carver, 1993 and Thompson, 1993), there appeared subsequent articles by Thompson (1996), Robinson and Levin (1997), and others, culminating in a debate on the topic at the April, 1998 annual meeting of the American Educational Research Association in San Diego.

The position taken here

This writer takes a very simple approach to the controversy. If there is a hypothesis to be tested and if a statistical inference is warranted (for a probability sample drawn from a well-defined population), then significance testing should be used. (The terms "hypothesis testing" and "significance testing" are regarded as interchangeable, but see Huberty, 1993 concerning the distinctions that are sometimes made between the two.) If there is no hypothesis to be tested but a statistical inference is warranted, then interval estimation (constructing a confidence interval around a point estimate) should be employed. If a statistical inference is not warranted (when the obtained data are for a full population or for a non-probability sample), whether or not there is a hypothesis to be tested, descriptive statistics should suffice.

One can often get hypothesis testing "for free" by using interval estimation (if the hypothesized parameter is not in the confidence interval, reject it), but there are situations where that is not the case (see Dixon and Massey, 1983, p. 93). When dealing with percentages, differences between percentages, or ratios of percentages, for example, the standard errors for the hypothesis-testing approach and the interval-estimation approach may differ considerably (see Knapp and Tam, 1997). For odds ratios associated with 2x2 contingency tables the significance test is straightforward, whereas the determination of the corresponding confidence interval is extremely complicated (see Fleiss, 1981, pp. 71-75).

Regression analysis

It is indeed curious that the adversaries in the significance testing controversy rarely use examples involving regression analysis (Steiger and Fouladi, 1997 is a notable

exception), which is the statistical technique that is most commonly used in the behavioral sciences.[2] There are many textbooks (e.g., Cohen and Cohen, 1983; Darlington, 1990; Marascuilo and Levin, 1983; Pedhazur, 1997; Stevens, 1996) and monographs (e.g., Achen, 1982; Berry, 1993; Berry and Feldman, 1985; Breen, 1996; Fox, 1991; Hardy, 1993; Iversen, 1991; Jaccard, Turrisi, and Wan, 1990; Jaccard and Wan, 1996; Langbein and Lichtman, 1978; Lewis-Beck, 1980; Newbold and Bos, 1985; Schroeder, Sjoquist, and Stephan, 1986)[3] that treat regression analysis. Hypothesis testing is given much greater emphasis than interval estimation in those sources. Most never even mention confidence intervals or devote very little space to their use (despite the fact that such intervals are routinely provided in the output of certain computer programs), suggesting that significance testing is the preferred approach. Of all of these authors, the only one who provides any sort of extended discussion of the advantages and disadvantages of confidence intervals vs. significance tests is Achen (1982), and he doesn't take a stand on one approach in preference to the other. Most users of regression analysis apparently are content with testing hypotheses concerning correlation coefficients (simple and multiple), regression coefficients (standardized or unstandardized), intercepts, and the like.

Some comments regarding observed power

There has recently been a disturbing tendency (disturbing to this writer and to a few others—see, for example, Goodman and Berlin, 1994, and Zumbo and Hubley, 1998) in some textbooks, journal articles, and computer programs to report the "observed power" for a study (see, for example, Munro, 1997 and the output for some of the analysis of variance programs in SPSS). Power is, or at least should be, an a priori concept. Researchers know (or should know), GOING INTO a study, the probability of getting a statistically significant finding (given the alternatively hypothesized effect size, the specified alpha level, and the sample size), i.e., the probability of rejecting a false null hypothesis in favor of a true alternative hypothesis. What some people are arguing for these days is the calculation of the obtained effect size (that's fine) and the determination of the corresponding "observed power" (that's not), COMING OUT OF a study. The rationale goes something like this: I'm willing to take the obtained **sample** effect size as a good estimate of the **population** effect size, see what power I had for that effect size for the sample size I drew, and determine what sample size I would need in my next study in order to have the power I want. That sort of reasoning seems terribly convoluted and an inappropriate use of power analysis as an aspect of statistical inference. Those who are interested in a counter-argument regarding the concept of "observed power" are urged to read the articles by Falk, Hogan, Muller, and Jennette (1992) and by Taylor and Muller (1995) and come to their own conclusions about the defensibility of that concept. The first of those articles is a substantive article concerning an experiment involving a fixed sample size (a priori power was not involved in its determination) of 26 people randomly assigned to two treatments, for which the research hypothesis is null, i.e., the theoretical position is that there is no treatment effect. (They found none and the study was terminated before the originally anticipated date.) The second article is a methodological article that advocates the calculation of obtained power for the Falk experiment for varying effect sizes close to null, and the construction of one-sided confidence intervals around those powers AND one-sided confidence intervals for the associated sample sizes.

Steiger and Fouladi

In defending their preference for interval estimation in multiple regression analysis (they also advocate the reporting of observed power), Steiger and Fouladi (1997) give the example of a confidence interval for the squared multiple correlation coefficient. The obtained R^2 in a sample of 45 observations on six variables (five independent and one dependent) was .40, which was statistically significant at the .001 level; the limits of the 95% confidence interval for the population R^2 were .095 and .562. They claim that the inference provided by the interval estimate is much more informative, albeit less impressive, than the inference provided by the significance test. That may be, but the price that was paid to get it (computationally complex calculations that are not included in standard statistical packages—but are available from Steiger and Fouladi) may not be worth it. This writer personally prefers the significance test, for a **given** null hypothesis, a **given** alternative hypothesis, a pre-specified alpha, and a sample size that is appropriate for a **given** desired power. Cohen's well-known and readily-available power book (Cohen, 1988) contains all of the necessary formulas and tables. There are also several readily-available software packages for carrying out such analyses.

Conclusion

This article has tried to summarize when significance tests (hypothesis tests) should be used and when they should not. Traditional regression analysis is one of the contexts in which tests of statistical significance appear to be most defensible and for which the corresponding interval estimation procedures are either not appropriate or are unnecessarily complicated.

It could be that many educational researchers are "closet Bayesians". They would like to be able to determine the probability that the null hypothesis is true, given the data, but in classical statistical inference that is not possible, so they must settle for the probability of getting the data (or something even more extreme), given that the null hypothesis is true (see Cohen, 1994). That's when they get frustrated and are prone to making all sorts of mistakes in interpreting significance tests. But the cure for this is not the abandonment of significance tests; the cure is to use them properly and interpret them properly OR to come out of the closet and become a Bayesian (see Pruzek, 1997 and Berger, Boukai, and Wang, 1997 regarding those alternatives).

Footnotes

[1] It might be argued that educational research is just like psychological research, sociological research, or research in any of the other social sciences, but many years ago Gowin (1972) claimed that it is (or at least should be) distinctive. Education is primarily interventionist. Our society doesn't have to develop various curricula, pay some teachers more than others, etc., but it has chosen to do so. It is therefore appropriate that controlled experiments and large correlational studies be carried out in order to determine to what extent such things "work".

[2] In their summary of statistical techniques used in reports of studies published recently in the *American Educational Research Journal*, the *Educational Researcher*, and the *Review of Educational Research*, Elmore and Woehlke (1998) indicated that multiple regression analysis was used in 148 out of 1906 articles (7.8 %), but if you add to that the 99 articles that used bivariate correlation, the 70 articles that used a t test, the 221 articles that used the analysis of variance or covariance (all of which can be subsumed under regression analysis—see, for example, Cohen, 1968) the total is 538 out of 1906 (28.2 %).

[3] These monographs were all categorized under the "Regression" grouping in a recent Sage University Paper.

References

Achen, C.H. (1982). *Interpreting and using regression*. Beverly Hills, CA: Sage.

Berger, J.O., Boukai, B., and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science*, *12*, 133-148.

Berry, W.D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.

Berry, W.D., and Feldman, S. (1985). *Multiple regression in practice*. Beverly Hills, CA: Sage.

Breen, R. (1996). *Regression models*. Thousand Oaks, CA: Sage.

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

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

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426-443.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd. ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

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

Cohen, J., and Cohen, P. (1983). *Applied regression/correlation analysis for the behavioral sciences* (2nd. ed.). Hillsdale, NJ: Erlbaum.

Darlington, R.B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Dixon, W.J., and Massey, F.J. (1983). *Introduction to statistical analysis* (4th ed.). New York: McGraw-Hill.

Elmore, P.B., and Woehlke, P.L. (April, 1998). Twenty years of research methods employed in *American Educational Research Journal*, *Educational Researcher*, and *Review of Educational Research*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Falk, R.J., Hogan, S.L., Muller, K.E., and Jennette, J.C. (1992). Treatment of progressive membranous glomerulopathy. *Annals of Internal Medicine*, *116*, 438-445.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd. ed.). New York: Wiley.

Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage.

Goodman, S.N., and Berlin, J.A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*, 200-206.

Gowin, D.B. (1972). Is educational research distinctive? In L.G. Thomas (Ed.), Philosophical redirection of educational research (Chapter I, pp. 9-25). The Seventy-first Yearbook of the National Society for the Study of Education. Chicago: The University of Chicago Press.

Hardy, M.A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (Eds.) (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, *61*, 317-333.

Iversen, G.R. (1991). *Contextual analysis*. Newbury Park, CA: Sage.

Jaccard, J., and Wan, C.K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.

Knapp, T.R., and Tam, H.P. (1997). Some cautions concerning inferences about proportions, differences between proportions, and quotients of proportions. *Mid-Western Educational Researcher*, *10*, 11-13.

Langbein, L.I., and Lichtman, A.J. (1978). *Ecological inference*. Beverly Hills, CA: Sage.

Levin, J.R. (1998). To test or not to test H_0. *Educational and Psychological Measurement*, *58*, 313-333.

Lewis-Beck, M.S. (1980). *Applied regression*. Beverly Hills, CA: Sage.

Marascuilo, L.A., and Levin, J.R. (1983). *Multivariate statistics for the social sciences*. Monterey, CA: Brooks/Cole.

Meehl, P.E. (1970). Theory testing in psychology and physics: A methodological paradox. In D.E. Morrison and R.E. Henkel, *The significance test controversy*. (Chapter 26, pp. 252-266). Chicago: Aldine.

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

Munro, B.H. (1997). *Statistical methods for health care research* (3rd. ed.). Philadelphia: Lippincott.

Newbold, P., and Bos, T. (1985). *Stochastic parameter regression analysis*. Beverly Hills, CA: Sage.

Pedhazur, E.J. (1982). *Multiple regression in behavioral research* (2nd. ed.). New York: Holt, Rinehart, and Winston.

Pruzek, R.M. (1997). An introduction to Bayesian inference and its applications. In L.L. Harlow, S.A. Mulaik, and J.H. Steiger (Eds.), *What if there were no significance tests*? (Chapter 11, pp. 287-318). Mahwah, NJ: Erlbaum.

Robinson, D.H., and Levin, J.R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*, 21-26.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Schroeder, L.D., Sjoquist, D.L., and Stephan, P.E. (1986). *Understanding regression analysis*. Newbury Park, CA: Sage.

Steiger, J.H., and Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik, and J.H. Steiger (Eds.), *What if there were no significance tests*? (Chapter 9, pp. 221-257). Mahwah, NJ: Erlbaum.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd. ed.). Hillsdale, NJ: Erlbaum.

Taylor, D.J., and Muller, K.E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *The American Statistician*, *49*, 43-47.

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*, 26-30.

Thompson, B. Review of What if there were no significance tests? *Educational and Psychological Measurement*, *58*, 334-346.

Zumbo, B.D., and Hubley, A.M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, *47* (Part 2), 385-388.