Comment

Contents

The Inscrutable Null Hypothesis

Warren W. Tryon
Fordham University

Hagen's (January 1997) article praising the null hypothesis statistical test (NHST) also cited literature critical of it and recognized that NHST "has been misinterpreted and misused for decades" (p. 22). NHST criticism goes back farther than the 30 years acknowledged by Hagen. Pearce (1992) reported that criticism of NHST began immediately with Fisher's introduction of it in 1925. Despite continuous critical commentary over the past 72 years, NHST became the primary method of data analysis in the social sciences.

A principal human factors requirement of any viable data analytic procedure, regardless of its other merits or demerits, is that it can be correctly calculated and interpreted. Widespread access to commercial statistical packages indicates that NHST calculations reported in the literature are probably correct. However, substantial reasons seriously question whether NHST results have been, are, or can be correctly interpreted consistently by most investigators.

Carver (1978) identified several misinterpretations of NHST results and reported that practices were unchanged 15 years later (Carver, 1993). Dar, Serlin, and Omer (1994) surveyed three decades of NHST misuse published in the *Journal of Consulting and Clinical Psychology* between 1967 and 1988. Cohen (1994) cited texts written by six prominent psychometricians that misinterpret NHST results. McMan (1995) found substantial NHST errors in most of 24 introductory psychology textbooks published between 1965 and 1994. Hagen's (1997) need to reprove three of Cohen's (1994) NHST criticisms indicates that even a prominent author of multiple statistics texts seemingly cannot "correctly" interpret NHST results. How much more susceptible to misinterpretation are the vast majority of other less well quantitatively trained psychologists?

Regardless of the technical merits or demerits of NHST, the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are undoubtedly substantial and may be the strongest reason for adopting other data analytic methods. Hagen's (1997) praise of NHST may be supportable on purely technical grounds but is unfortunate if it prolongs primary reliance on NHST to evaluate quantitative difference and equivalence given the prominent human factors problem of widespread and intractable interpretation errors. Alternative methods are available for these purposes that are far less subject to misinterpretation. The science of psychology can only benefit by supplementing, if not replacing, NHST practices with these methods.

REFERENCES

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

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75–82.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

McMan, J. C. (1995, August). *Statistical significance testing fantasies in introductory psychology textbooks.* Paper presented at the 103rd Annual Convention of the American Psychological Association, New York.

Pearce, S. C. (1992). Introduction to Fisher (1925): Statistical methods for research workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics, Vol. 2: Methodology and distributions* (pp. 59–65). New York: Springer-Verlag.

Correspondence concerning this comment should be addressed to Warren W. Tryon, Department of Psychology, Fordham University, Bronx, NY 10458-5198. Electronic mail may be sent to wtryon@murray.fordham.edu.

Significance Testing: Is There Something Better?

Robert E. McGrath
Fairleigh Dickinson University

In the article "In Praise of the Null Hypothesis Statistical Test" (NHST), Hagen (January 1997) did an admirable job of reminding readers that NHST represents a brilliant and useful innovation, with relevance for research settings far more sophisticated than those originally considered by its creator. However, it is important to note that even this very supportive article does not offer a strong case for its continued use as the primary inferential strategy in psychology. I address five particular aspects of the article.

First, Hagen (1997) suggested that "if we are content to equate the $P(H_0)$ with a subjective degree of belief, or level of confidence, then the NHST does, indeed, tell us what we want to know" (p. 19). Instead, what Hagen demonstrated is what a Bayesian analysis of NHST results can reveal about

 0003-066X/98/$2.00
Vol. 53, No. 7, 796–807

$P(H_0)$. This is not a trivial distinction. Bayes's theorem is rarely taught to psychologists or used as an adjunct to NHST. This is in part the fault of Fisher (1937) himself, who specifically opposed the use of Bayes's theorem in this context. It is worth noting that the same argument raised by Hagen in support of NHST was originally introduced by critics of the method who took Fisher at his word about how NHST was supposed to proceed (see Oakes, 1986).

Second, Hagen (1997) responded to the popular belief that an effect size exactly equal to zero is unlikely, rendering an analysis aimed at evaluating whether the effect equals zero absurd. His argument seems to be that although in any one study the effect will never equal zero, there is no reason to believe these discrepancies will not even out across studies, leaving the null hypothesis true at the level of the population.

Although it is true as noted that a zero effect is not impossible, it is highly unlikely that an effect will exactly equal zero in anything less than the most well-controlled studies. As noted by Cohen (1994), there is no reason to expect that population correlations between uncontrolled variables exactly equal zero, so the use of no-effect significance tests in any observational study is suspect. Even in well-controlled true experiments, there are often nonrandom nuisance variables inherent to the experimental design that cannot be perfectly controlled (Campbell & Stanley, 1963). Hagen (1997) himself stated that "Tukey's (1991) comment that the effects of A and B are always different can stand. But it does not necessarily follow that the null hypothesis will always be vulnerable to those effects" (p. 21). This is a far cry from saying that the null hypothesis can usually be considered invulnerable to these effects, a statement that would be more consistent with recommending the widespread use of NHST.

Third, NHST has been criticized because as a system for the testing of propositions, it does not demonstrate the same level of logical validity as the modus tollens. Hagen (1997) accepted this critique but responded by suggesting that evaluations of scientific propositions rarely demonstrate the highest level of logical validity. It is an interesting argument but again begs the question of whether there are more logically justifiable methodologies.

Fourth, Hagen (1997) responded to Cohen's (1994) and Schmidt's (1996) preference for confidence intervals over NHST by suggesting that confidence intervals are no better than NHST for the purpose of testing null hypotheses. This is true, but it ignores the primary reason for preferring confidence intervals. Under NHST, the basic question in primary research is "Based on this sample, what is our best guess about whether or not ρ equals 0?" The computation of confidence intervals allows for a much more interesting question: "Based on this sample, what is our best guess about the value of ρ?" This represents a fundamental change in the way that the analytic process is conceptualized. It is only if one fails to look beyond the limits of NHST that confidence intervals and NHST appear to be equivalent strategies.

Finally, Cohen (1994) clearly did not intend his article to be a comprehensive review of the problems associated with significance testing. By narrowly focusing on the arguments raised by Cohen, Hagen (1997) ignored the bulk of the criticisms leveled against the method. These criticisms include, among others, the logical problems associated with making a binary decision, the inevitably arbitrary element in the selection of alpha, the negligence of sample size issues fostered by Fisher's (1937) model of NHST, and obstacles to the accumulation of knowledge in psychology created by the use of NHST (Schmidt, 1996). Hagen's conclusion that "I have tried to point out . . . that the logic underlying statistical significance testing has not yet been successfully challenged" (p. 22) seems particularly excessive given the limited range of his response.

Hagen (1997) as well as Frick (1996) offered good, albeit incomplete, responses to those who would suggest NHST is useless or hopelessly, logically flawed. However, I do not think the question has ever really been "Is it useless?" but rather "Is there something better?" This question deserves much closer scrutiny than is possible here, but a popular opinion holds that interval estimation represents a superior strategy to NHST in many ways. Given all that has been gained through its use, I think it is very appropriate to praise the brilliance of NHST, but having done so, perhaps it is time to bury it.

REFERENCES

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

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

Fisher, R. A. (1937). *The design of experiments.* London: Oliver & Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester, England: Wiley.

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

Correspondence concerning this comment should be addressed to Robert E. McGrath, School of Psychology, T110A, Fairleigh Dickinson University, Teaneck, NJ 07666. Electronic mail may be sent to mcgrath@alpha.fdu.edu.

In Praise of Value Judgments in Null Hypothesis Testing . . . and of "Accepting" the Null Hypothesis

Robert G. Malgady
New York University

As Hagen (January 1997) acknowledged and as Cohen (1994) did before him, there has been considerable discourse on the merits and limitations of null hypothesis testing, dating back to Ronald Fisher (1935) himself. Nonetheless, as insightful as even Hagen's illumination of null hypothesis testing is, I believe two related issues have been obscured, if not neglected.

As most statisticians and philosophers of logic would say, one can reject a null hypothesis but can never accept or validate it by using the classical Fisherian (Fisher, 1935) procedure. I have argued elsewhere that in clinical research, this is fundamentally like burying one's head in the sand (Malgady, 1996). If the null hypothesis is not rejected in a statistical test, one certainly cannot assert that it has scientific validity, but behavior concerning the null hypothesis validates it because people act as if it were true. For instance, if a psychopharmacological researcher tests a new drug for treating major depression disorder, a null hypothesis might be that mean reduction of depressive symptomatology does not differ between an experimental (drug) condition and a placebo control. If this null hypothesis is not rejected, the researcher cannot lay scientific claim to its validity. But the obvious consequence of this decision is that, rightly or wrongly so, the drug will not be prescribed for persons with major depression disorder. Science dictates a conservative or skeptical stance—scientists don't believe in something until there is evidence of its truth within, of course, a comfortable margin of risk (e.g., probability of being in error < .05). Thus, there is a family of status quo null hypotheses composing a