The Behavior Analyst 1999, 22, 87–92 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Some Things Significance Testing Does and Does Not Do

Marc N. Branch
University of Florida

Significance testing plays a prominent role in behavioral science, but its value is frequently overestimated. It does not estimate the reliability of a finding, it does not yield a probability that results are due to chance, nor does it usually answer an important question. In behavioral science it can limit the reasons for doing experiments, reduce scientific responsibility, and emphasize population parameters at the expense of behavior. It can, and usually does, lead to a poor approach to theory testing, and it can also, in behavior-analytic experiments, discount reliability of data. At best, statistical significance is an ancillary aspect of a set of data, and therefore should play a relatively minor role in advancing a science of behavior.

One need only look at a few scientific journals in the domain of behavioral science to note the ubiquity of statistical significance testing (cf. Hubbard, Parsa, & Luthy, 1997; Sterling, 1959). Such information can be found even in journals that emphasize replication and individual-subject data (Hopkins, Cole, & Mason, 1998). It is interesting that behavioral science has aligned itself so strongly with significance testing, whereas other, more successful sciences have not. Significance testing has been ubiquitous in psychology since the early 1950s, yet it is difficult to discern how its use has improved the field. Many important behavioral processes were discovered and analyzed prior to the development and implementation of tests of statistical significance, and virtually all of modern chemistry and physics was developed without their assistance. It is clear, then, that science can proceed without significance testing, and it is not at all clear that significance testing has helped to advance behavioral science. The mystery is why so many behavioral scientists continue to place such high value on the methods of significance testing.

Many journals require tests of statistical significance as part of the data analyses. Given that it is possible to conduct good research without resorting to this method, it seems unwise for journals to have editorial policies that make such tests necessary for publication. Nevertheless, such policies exist, and that presents a problem for the researcher who prefers not to employ the tests. In the comments that follow, I hope to help researchers who find themselves at the mercy of journal editors or grant reviewers who clamor for significance tests use the old maxim, "Sometimes the best defense is a good offense." Specifically, I'll try to point to both inherent weaknesses in the logic of significance testing and to potential consequences of unreasonable allegiance to them. These points may be used in interchanges with editors and reviewers to illustrate that preference for not using significance testing is fully defensible. Virtually none of the weaknesses and negative consequences of significance testing that I shall present are new or recently discovered. All have been known and discussed many times over the years. My comments, then, should stand only as reminders.

Preparation of the paper was assisted by USPHS Grant DA-04074 from the National Institute on Drug Abuse.

Reprints can be obtained from the author at the Psychology Department, University of Florida, Gainesville, Florida 32611 (E-mail: branch@psych.ufl.edu).

THINGS THAT SIGNIFICANCE TESTING DOES NOT DO

Tests of Statistical Significance Do Not Provide a Quantitative Estimate of the Reliability of a Result

This fact means that the expression "statistically reliable" is a non sequitur. There is no dispute about this issue. There is no known relationship between level of significance and replicability (see Carver, 1978). A p value, or its inverse, therefore is not an estimate of how likely the results obtained are to be replicated. That is, to say that $p < .001$ does not imply that there is only one chance in a thousand that a replication would fail, nor does it mean if you conducted the experiment 1,000 times, you would most likely find only one discrepant result. The only way currently known to determine if a finding is reliable is to replicate the result. How many replications must be performed before we agree that a result is reliable? The answer is, "It depends." And it depends on many things. The number of replications needed to be convincing depends, for example, on what is already known. As an illustration, if I throw a brick at a pane of glass and it breaks, most people would not ask that I replicate the effect to be sure to conclude that the brick hitting the glass was the cause of the glass shattering. (Nor would most people ask for a "control" pane of glass, but that is another, although related, issue.) Why not? Because of what we already know about glass, bricks, and thrown objects. Similarly, in science, some things, because we know relatively little, require more extensive replication, whereas other kinds of results do not require as much (these are more likely in physics, however, than in behavioral science).

A p value has quantitative meaning only if the null hypothesis is true. Of course, we do not know if the null hypothesis is true. That is why we're doing the test in the first place—to get information to help us decide if it is true. If we knew that the null hypothesis were true, then a p value of .001 would indicate that if we perfectly replicated the study 1,000 times, we should expect only one case to come out differently. Of course, if we *knew* the null hypothesis were true, we would have no reason to do the test.

Tests of Statistical Significance Do Not Estimate the Probability That the Results Were Due to Chance

This fact can be illustrated easily by remembering that a p value is a conditional probability. Specifically, a p value represents the probability of a certain kind of data given that the null hypothesis is true: $p = p(\text{data}|H_0)$. To state that something is due to chance is to reverse the conditionality. That is, saying that some result is due to chance is essentially stating that a p value is an estimate that the null hypothesis was operating, and that simply is not the case because that would imply that $p = p(H_0|\text{data})$. Elementary probability theory tells us that $p(A|B)$ is not equal to $p(B|A)$, except in the rare case that the probabilities are independent of one another and are also equal (Parzen, 1960). It is easy to illustrate to oneself that they are not equal by considering everyday examples like p(raining|cloudy) versus p(cloudy|raining) or p(manuscript rejected|submitted to *The Behavior Analyst*) versus p(submitted to the *The Behavior Analyst*|manuscript rejected). All this is to say that p is not an estimate of the probability of the truth of the null hypothesis. That is one of the important reasons why it is not an estimate of the reliability of results.

Tests of Statistical Significance Usually Do Not Answer a Question to Which the Answer Is Unknown

Significance tests provide researchers with evidence on which to decide if the null hypothesis is true. That is not a very important question to answer, because in the vast majority of cases the null hypothesis of no differ-

ence is not true. As Meehl (1967) notes, he and Lykken have shown that to be the case with data. When one considers what the null hypothesis usually entails (i.e., that the only thing operating in the experiment is randomness), it is pretty obvious that there are very few situations in which that could be true. As noted by Kraemer (1998), "something nonrandom is almost always going on, and it seems a trivial exercise to redemonstrate that fact" (p. 206). Virtually all behavioral scientists learn that by increasing *N* one increases the likelihood of finding statistical significance, but somehow the implications of that fact get lost. If it is so important, how come it is so easy to influence? One can relate this point to the previous two. Perhaps it is not so bad that significance tests do not estimate the truth of the null hypothesis, because we already know that it is false. Also, a procedure designed to help us decide about something we already know could hardly be one that would provide quantitative estimation of reliability.

It seems to me that the three arguments just presented would be sufficient to convince a reasonable person that eschewing significance testing for other approaches to establishing reliability is a perfectly sane and defensible position. Perhaps, however, one might be faced with someone who is less than reasonable and therefore might want to use the following points about unfortunate consequences of null-hypothesis significance testing.

THINGS THAT SIGNIFICANCE TESTING DOES DO

Tests of Statistical Significance Reduce Scientific Responsibility

This point is made eloquently by Carver (1978), who points out that slavish adherence to significance testing is a view in which a scientist is given full responsibility for the origin, design, and conduct of an experiment, but is given no responsibility for deciding whether the results are useful or meaningful. The thing that sets science apart as a social enterprise is that it is self-corrective. The mechanism of correction is replication. It is through replication that confidence in a finding is established, and it is through failures of replication that mistakes are corrected. Social contingencies are important in science, and significance testing blunts their effectiveness. Before tests of significance were invented, a scientist's reputation depended on the reliability of his or her descriptions of results and conclusions drawn from them. If a scientist claimed to observe some result, and subsequently it was shown that the result was not reliable, the scientist's reputation suffered. With significance testing, there is an out. For example, suppose Scientist A performs an experiment and gets a statistically significant result and makes claims on that basis. Other scientists perform replications but do not get the same result. Does Scientist A's reputation suffer? No, because he or she can claim, "It's not my fault. We expect some proportion of errors when using tests of statistical significance, and this was one of them. I played by the rules and am therefore blameless." If one's reputation rides on the reliability of findings, you can bet that scientists would be more careful about what they publish.

The fact that tests of statistical significance protect a scientist to some degree may be one of the factors that determine their popularity. Use of significance testing might be thought of as a form of avoidance, avoidance of social censure.

Requiring statistical significance as a prerequisite for publication also serves to blunt science's most precious resource for ferreting out mistakes (or even fraud). As noted above, the self-corrective nature of science is based on replication. Failures to replicate are very, very important in informing us that we don't fully understand what is going on (Sidman, 1960). If statistical significance is requisite for publication, then one will have a difficult time pub-

lishing results of experiments in which replication of a statistically significant effect fails.

Tests of Statistical Significance Are Frequently Employed in a Poor Manner to Test Theory

I like to call this the "dumb-null-hypothesis problem," and it is a point well made by Meehl (1967, 1978). Consider the usual arrangement for using statistical significance testing to put a theory to test. The null hypothesis is set at "no effect." The alternative hypothesis, the one that the theory predicts, is set at "some effect." If a statistically significant effect is obtained, the null hypothesis is rejected, and the alternative hypothesis, and therefore the theory, gain support. Consider now how statistical significance is determined. A statistic that is a ratio of "effect variance" over "error variance" is computed. If that ratio is large enough, statistical significance is achieved. Next, consider the effects of improvement in experimental technique. Better control of extraneous variables should decrease error variance, and therefore make it easier for the ratio to reach the critical value. Thus, as methods are refined and better experiments are conducted, it becomes easier to demonstrate statistical significance. That means that better methods make it easier to reject the null hypothesis and therefore support the theory, no matter what the theory is. Obviously, this is not a very good outcome.

There is a way to circumvent this issue and make use of significance testing in a more rational fashion. Statistics courses usually inform us that it is not necessary that the null hypothesis be "no effect." Instead it can be set at some particular effect; that is, it can be set at what the theory predicts. Then, as experimental techniques are improved it still becomes easier to reject the null hypothesis, but in this case the null hypothesis is what the theory predicts. Therefore, as experimental techniques improve, the theory is put to a more severe test, exactly the kind of result for which one would hope. This latter approach is exactly what is done in curve fitting (Lewis, 1966), a strategy favored by the more advanced sciences.

Tests of Statistical Significance Emphasize Population Parameters Over Behavior

As Danziger (1987, 1990) has noted, a remarkable development in behavioral science in the last half of this century has been the emergence of the aggregate as the unit of analysis. This is an odd development in a field presumably dedicated to understanding behavior or "the mind." Behavior is something an individual does, not what a group average does. (It is especially difficult to think of "group mind.") The direction of inference from a group average is to the population, not to the individual, so when the unit of analysis becomes the aggregate we develop a science not of behavior but of population parameters. Perhaps some take comfort from the view that something that provides information about the population is inherently more general than something that applies to some individuals, but that comfort ought to be tempered by the realization that the generality is not about behavior. A 2% rate of pregnancy in a population may have important meaning for policy makers (insurance and public), but it has no meaning for an individual female, who is never 2% pregnant.

Tests of Statistical Significance Can Limit the Reasons for Doing Experiments

Tests of statistical significance generally require that the experimental question be a test of a hypothesis. Testing hypotheses, of course, is an honored tradition in science and certainly a worthy enterprise. As ably noted by Sidman (1960), however, there are many other very good reasons for doing experiments. Isaac Newton, a sci-

entist of some note and success, suggested that one should never have a hypothesis, but rather should simply ask questions about Nature. His successes make clear that hypothesis testing is not the only route to achievement in science. Having to shoehorn one's experiments into the logic of hypothesis testing frequently leads to absurdities like developing one's "hypotheses" after the data are collected.

Tests of Statistical Significance Often Discount Reliability in Effects in the Case of Behavior-Analytic Experiments

This happens because the tests typically ignore the replications inherent in behavior-analytic research designs. Consider, for example, cases in which for each subject a stable baseline is established in each condition of the experiment (a very common occurrence in such research). Statistical analyses will then proceed using, for example, averages from the last five sessions of observation in each condition. This mean is treated as if it were a single score from a single observation, but it clearly is not. Each of the last five sessions of each condition constitutes a replication, so in reality, at a minimum, five replications of the value are ignored. Given other evidence of good experimental control, five replications of a value provide direct information about the reliability of the value, and that information is lost in the statistical test. The expression, "at a minimum," was used above because in many cases the session average itself underestimates the reliability of the measure. Suppose that in each session a variable-interval schedule was in effect, and cumulative records reveal that rate of behavior was constant throughout the session. If the rate of behavior is reported as the session average, this single number does not tell us as much about reliability of the effect. A cumulative record, however, reveals that from minute to minute the effects were reliable. I know of no statistical test that can deal with that kind of reliability.

Given that tests of statistical significance, despite any evidence that they have assisted the development of behavioral science, have become such an integral feature of research in behavioral science, it seems highly unlikely that we shall at any time soon see a broad deemphasis of their use. There are rumblings, however, in the psychological sciences that may indicate that slavish attachment to significance testing may eventually fade away (e.g., Cohen, 1994; Hunter, 1997; Loftus, 1996). If that is the case, research conducted in the tradition of behavior analysis, research that is directed at individual behaving subjects and that employs methods that directly illustrate reliability, can serve as a model for other researchers. Now could well be a very opportune time for behavior analysis, a time in which behavior analysis illuminates the way for other researchers in psychology. Behavior analysts should not "hitch themselves to a wagon from which people are leaping," but instead should look upon the next decade as one in which behavior-analytic methods gain even wider popularity.

REFERENCES

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

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

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Kruger, G. Gigerenzer, & M. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 35–47). Boston: MIT Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* New York: Cambridge University Press.

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21,* 125–137.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the *Journal of Applied Psychology,* 1917–1994. *Theory and Psychology, 7,* 545–554.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 307.

Kraemer, H. C. (1998). Statistical significance: A statistician's view. *Behavioral and Brain Sciences, 21,* 206–207.

Lewis, D. (1966). *Quantitative methods in psychology.* Iowa City: University of Iowa Press.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Parzen, E. (1960). *Modern probability theory and its applications.* New York: Wiley.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association, 54,* 30–34.