If you've got an effect, test its significance; if you've got a weak effect, do a meta-analysis

John F. Kihlstrom

Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650. **kihlstrm@cogsci.berkeley.edu socrates.berkeley.edu/~kihlstrom**

Abstract: Statistical significance testing has its problems, but so do the alternatives that are proposed; and the alternatives may be both more cumbersome and less informative. Significance tests remain legitimate aspects of the rhetoric of scientific persuasion.

I admit it: after more than 25 years of reading, writing, reviewing, and editing scientific research in psychology and related fields, I still cannot understand the fury that whirls around statistical significance testing. Yet the critics seem to be gaining ground: the *American Journal of Public Health* virtually banned tests of statistical significance from its pages, at least for a time, and the American Psychological Association (APA) has seriously contemplated doing the same. Whatever the outcome of the APA's deliberations, the pages of *Psychological Science,* the flagship journal of the American Psychological Society, will remain open to significance tests so long as I remain editor. The reasoning behind this policy is more pragmatic than mathematical, but I am glad to have my view bolstered by Chow's (1996) cogent, scholarly analysis of the debate.

Criticisms of significance testing, at least within psychology, take two broad forms (for representative samples of these criticisms, see Gonzalez 1994; Hunter 1997; Loftus 1996; Schmidt 1996; for responses to Hunter's paper, see Abelson 1997; Estes 1997; Harris 1997; Scarr 1997; and Shrout 1997). On the one hand, it is argued that when the sample size is large enough, even trivial effects can achieve statistical significance. Thus, effects can be touted as "significant" that are in fact utterly trivial from the standpoint of either theory or practice. On the other hand, it is argued that the failure to achieve statistical significance causes investigators (and other consumers of research) to discount effects that might well be of theoretical interest or practical importance. Thus, significance tests either deliver too much, by portraying negligible effects as consequential, or too little, by insinuating that genuine effects are nonexistent.

Rather than test for statistical significance, researchers are sometimes advised to report confidence intervals instead. But confidence intervals only make sense when the goal of the research is to make a point estimate – for example, of the mean family income for African Americans, or how many people will vote Republican in the next election. In such cases, it is ridiculous to test the null hypothesis, and researchers are well advised to calculate confidence intervals as an index of the precision of their estimates. But psychologists rarely wish to estimate population parameters; rather, we generally test hypotheses about the effects of particular treatments (e.g., two levels of distraction on memory), or about the relations between particular variables (e.g., two dimensions of personality), which have been manipulated or assessed because they are theoretically or practically interesting.

Suppose, for example, that a researcher publishes a study in which psychiatric patients who receive imipramine score, on average, 5 points lower on a depression scale than those who do not, whereas the difference averages 10 points for those who receive fluoxetine. Should a researcher simply report these point estimates? Certainly not, because point estimates cannot speak for themselves. In the first place, we're not interested in the point estimates, because they would be entirely different if the researcher had used a depression test with different scaling properties. What we really want to know is: do either of these effects differ from what would be observed in a placebo group? Do any of these effects differ from zero? And do any of these effects differ from each other?

These questions can be answered by calculating the confidence intervals around each mean, and then determining the extent to which these intervals overlap. But isn't it much easier on everyone if the researcher simply reports the results of an analysis of variance followed by planned comparisons, adopting a conventional level of statistical significance like $p < .05$ or .01? It is important to bear in mind, as Chow (1996) clearly demonstrates, that comparing confidence intervals and testing statistical significance are, for all intents and purposes, mathematically equivalent (remember the debate over analysis of variance versus multiple regression?). And significance tests give you a p value to boot!

Of course, in this instance, significance testing might well indicate that neither of the drugs differs from placebo and that none of the means differ either from the others or from zero. Now suppose that a dozen more such studies are published, each yielding null results, but that a meta-analysis of the baker's dozen shows that, in fact, the effects of fluoxetine are greater than those of imipramine, which in turn are greater than those of placebo, which in turn are greater than zero. In this case, it is true that the failure of the first study to reject the null hypothesis is misleading: fluoxetine and imipramine are better than nothing. But the problem does not lie in statistical significance testing; rather, it lies in the researchers' failure to perform studies with enough power to reject the null hypothesis in the first place, the reviewers' failure to detect this flaw, the editor's willingness to accept the papers for publication, and the readers' willingness to take them seriously.

Even if the initial study had yielded significant results, of course, there might have been problems. With huge *N*s, even trivial differences can achieve statistical significance. So, investigators and consumers of research alike always have to ask themselves whether they should really care about a "statistically significant" result. How much variance is accounted for by the effect? Reporting effect sizes helps in this assessment, but in the final analysis the standards for small, medium, and large effects (Cohen 1992) are no less arbitrary (and no less context-specific) than the standards for statistical significance. In any event, it should be understood that none of these alternative techniques – statistical significance testing, comparison of confidence intervals, or meta-analysis – has any privileged status with respect to another important question: Are any of the treatment effects clinically significant (Jacobson & Christensen 1996; Jacobson & Truax 1991; Jacobson et al. 1984)? Clinical significance is sometimes assessed in terms of something like effect size, although it is not clear that the simple expedient of adopting stricter criteria for statistical significance would not yield the same conclusions. In the final analysis, however, the problem of clinical significance concerns the criteria by which treatment outcome is assessed rather than the statistical tools by which significance is documented.

I have dwelt on an example drawn from clinical research, but it should be clear that similar considerations apply to basic, theory-oriented research as well. Theories (formal or informal) generate hypotheses about the effects of certain manipulations, or the relations among certain variables, and statistical significance is often the most convenient way of testing these hypotheses. Chow (1996) does us a great service by pointing out that confidence intervals and effect sizes have little to offer when we wish to corroborate a scientific theory, where the hypotheses at stake are not at the same level of abstraction as "H_0 = P does not exist, H_1 = P does exist" – and I wish he had said more about Fisher's own role in the mistaken equation of significance testing with null hypothesis significance testing. Estes (1997) likewise reminds us that tests of statistical significance are the chief means of testing how well mathematical models or computer simulations of mental processes fit actual empirical data. Given that theory testing is the goal of science, and that formalisms such as operating computer simulations represent psychological theorizing at its best (Simon 1969), it would seem foolhardy to abandon statistical significance testing – even for those, like myself, whose theorizing never gets beyond the vague and verbal.

Significance tests are not our only means of analyzing and interpreting data, though, and we probably do rely too heavily on them. That statistical significance testing has become something of a fetish is indicated by the reflexive way in which many researchers (and not just novices) report artificially precise values (e.g., $p < .0438$) ripped from their computer printouts, instead of adopting conventional (and more conservative) ranges like .05, .01, .005, and .001); by their persisting tendency to report one-tailed tests when two-tailed ones would do just fine; and by their inclination to conclude that $p < .01$ is "more significant" than $p < .05$). While I am grateful for Chow's (1966) mathematical exegesis, I wish that he had said more about these sorts of practical matters.

In the final analysis, the value of significance testing is practical, as a component of the rhetoric of science (Abelson 1995). Researchers can have their own subjective opinions about their own and others' results, but statistical significance tests are – how else to put it? – public, empirical, *tests of significance.* They constitute a principled way for researchers to claim that their experimental results are worth knowing about, and for consumers to evaluate researchers' claims. At least since the time of Neyman and Pearson (1928) and Fisher (1935), significance testing has kept the behavioral, cognitive, and social sciences from lapsing into solipsism, and they can continue to play this role, along with all the other procedures in our statistical repertoire.

Statistical significance: A statistician's view

Helena Chmura Kraemer

Department of Psychiatry and Behavioral Science, Stanford University, Stanford, CA 94305. **hck@leland.stanford.edu**

Abstract: From a statistician's viewpoint, the concepts discussed by Chow relating to "statistical" significance bear little resemblance to the concept developed in statistics. Whether or not "statistical significance" has a place in psychological research is a decision for psychologists, not statisticians, to make, but the decision should be based on a less flawed version of what is being considered.

I generally agree with Chow's conclusion but disagree with much of his book. My objections would be allayed, however, were Chow to rename his book something like "Psychological significance" and to point out that his concepts had but a tangential relationship to statistical significance testing as developed in the field of statistics.

As Chow states, every research project begins with a substantive hypothesis. To establish its truth usually requires a convergence of evidence from many approaches, with null-hypothesis significance-testing procedures (NHSTP) but one of the many to be considered when the psychologists claim:

> *Claim: If I prove, beyond reasonable doubt, that "such and so" is true, the credibility of the substantive hypothesis will increase.*

The key phrase here is "beyond reasonable doubt," encapsulated in NHST in the significance level, α. Significance level is not a probability, conditional or otherwise. It is a number between 0 and 1 selected by the proponent as the *proposed upper limit* of the probability of any false claim that "such and so" is true. As such, it reflects (1) the scientific standards of the proponent and (2) what is acceptable to peer reviewers. What is so holy about $\alpha = 5\%$ or 1%? Nothing. Why can't it be 6% or 10% or 20%? However, α must be set *before* the evidence (data) is collected and analyzed and (2) peer reviewers must accept the levels as appropriate in the field of application.

Critical also is translating "such and so" into the "null hypothesis" H_0, which has two components: H_0: "such and so" is not true, and certain assumptions are true. Only if one proposes to demonstrate merely that "something nonrandom is going on" does the null hypothesis posit chance or randomness. As others have so eloquently pointed out, something nonrandom is almost always going on, and it seems a trivial exercise to redemonstrate that fact. Chow appears to believe that every H_0 posits randomness.

Moreover, certain design or mathematical assumptions are always incorporated into H_0 – assumptions with which the psychologists and statisticians (often tacitly) agree: normal distributions, equal variances, linear associations, and so forth, which play a key role in NHSTP.

> *Problem 1: Chow's definitions of "significance level" and "null hypothesis" are either incomplete or imprecise.*

In NHSTP the psychologists propose a research design that is to produce certain data, and the statisticians propose that H_0 be rejected when a selected test statistic falls into a specified region. To show that this is a valid α-level test, the statistician must show mathematically that *whenever* the null hypothesis is true, using this design and this proposed test, the probability of rejecting H_0 by the proposed rule never exceeds α.

> *Problem 2: Under the null, as well as under the non-null hypothesis, there are typically many distributions, one for each possibility. The appropriate graphic is the operating characteristic curve, a graph of the probability of rejecting H_0 when each such possibility is true, and not any one or two "bell-shaped" distributions.*

In any case it is uncommon that the distributions on which probabilities are based are exactly "bell-shaped" at all, but that's a quibble.

> *Problem 3: The formulation of the proposed testing procedure specifically depends on the form of the alternative hypothesis.*

To take the simplest example, the difference between a proposal for a one-tailed versus two-tailed *t*-test depends strictly on the formulation of the alternative hypothesis. Generally, selecting a NHSTP sensitive to the researcher's specific claim is an essential part of the process of selecting an appropriate NHSTP.

> *Problem 4: For any claim, there are many different valid NHSTPs, among which a choice must be made. Chow removes the primary basis for such a choice when he recommends against power analysis.*

Again, take the very simplest situation of a two-sample *t*-test: What is the proposed total sample size? What proportion of the total sample will be assigned to or selected from the two groups? Will the sample be stratified or matched? Will there be only an endpoint observation for each subject? If more, when and how? Is the two-sample *t*-test the best choice of test? Each of these decisions changes the NHSTP. Statisticians would base the choice on comparisons of power. How would Chow choose among them?

When researchers either reject or do not reject H_0 what is it that is rejected or not rejected? If one rejects H_0, what one logically accepts is Not-H_0: *Either the claim is true OR some of the assumptions are not true.* Any power calculations were done for the so-called alternative hypothesis: *The claim is true AND all the assumptions are true.* But the psychologists' claim was: *The claim is true.*

How well Not-H_0 or the alternative hypothesis corresponds to the claim depends on how well those assumptions correspond to reality. If crucial assumptions are not reasonably well satisfied, what does it matter whether or not results led to rejection of H_0? The results are likely invalid.

> *Problem 6: If the NHSTP is valid, rejecting H_0 or not rejecting H_0 is a comment on the strength of the evidence to make a certain claim, not a comment on the truth or falsehood of either the null or alternative hypotheses, on the effect size, or on the future replicability or confirmability of the conclusion. Each of these interpretations is at some time indicated by Chow's presentation.*

When one validly rejects H_0, one in effect says, "The evidence is strong enough to risk making a claim that 'such and so' is indeed true." When the result is "non-significant," one says: "The evidence is not strong enough to risk making any claim with regard to