The Null Hypothesis Testing Controversy in Psychology

David H. KRANTZ

A controversy concerning the usefulness of "null" hypothesis tests in scientific inference has continued in articles within psychology since 1960 and has recently come to a head, with serious proposals offered for a test ban or something close to it. This article sketches some of the views of statistical theory and practice among different groups of psychologists, reviews a recent book offering multiple perspectives on null hypothesis tests, and argues that the debate within psychology is a symptom of serious incompleteness in the foundations of statistics.

KEY WORDS: Foundations of statistics; Hypothesis tests; Psychometrics.

1. INTRODUCTION

This article began as a review of a recent book, *What If There Were No Significance Tests?*, edited by Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger. The book was edited and written by psychologists, and its title was well designed to be shocking to most psychologists. The difficulty in reviewing it for *JASA* is that the issue debated may seem rather trivial to many statisticians. The very existence of two divergent groups of experts, one group who view this issue as vitally important and one who might regard it as trivial, seemed to me an important aspect of modern statistical practice. I decided to discuss this divergence in my review, which subsequently grew (with kind-hearted editorial guidance) into an article.

The article is organized as follows. In Sections 2–4 I briefly discuss the nature of statistical expertise within psychology, the abuse of null hypothesis testing, and my view that even seemingly trivial difficulties in the practice of statistics deserve careful analysis. In Section 5 I sketch some of the different uses of hypothesis testing that need to be distinguished in a foundational analysis. I comment on the book in Section 6, and give a short conclusion in Section 7.

Although the book contains many perspectives and ideas, its central question concerns the usefulness of null hypothesis tests as a scientific research tool in psychology. This is sometimes viewed as a logical or philosophical question: To what extent does the logic underlying null hypothesis testing match the logic of scientific inference (in psychology or in other disciplines)? It can also be viewed as an empirical question: Is there a causal relation between the use of this tool and rapid or slow progress in psychological research?

This issue has been debated for about 40 years in psychological journals. One of the contributors to this volume was among the earliest critics of null hypothesis testing in psychology (Rozeboom 1960). In the last few years, however, the debate has become intense. Chapter 3, by Frank L. Schmidt and John E. Hunter, advocates an outright ban of significance tests, replacing them with point estimates accompanied by confidence intervals. Arguments against such a ban are offered in Chapter 4, by Stanley A. Mulaik, Nambury S. Raju, and Richard A. Harshman, and in Chapter 5, by "A Retrospective on the Significance Test Ban of 1999," by Robert P. Abelson.

Some readers of *JASA* may find this core issue too trivial to merit serious discussion. Because a point estimate, together with a t statistic for a particular hypothesized parameter value, can be readily converted to an (approximate) confidence interval, and vice-versa, the choice between these two forms of presentation might be considered a matter of style rather than substance. Analogously, some statisticians might prefer to drop the term "Fisher Z transform" in favor of "inverse hyperbolic tangent," and one could perhaps discuss the merits of such a stylistic change for 15 minutes or so, but writing a book on this issue would be absurd. Yet Schmidt and Hunter (and other test-ban advocates) are in dead earnest. Their abstract for Chapter 3 concludes with the following passage:

> Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution. After decades of unsuccessful efforts, it now appears possible that reform of data analysis procedures will finally succeed. If so, a major impediment to the advance of scientific knowledge will have been removed.

Similar statements have appeared elsewhere recently. For example, a prominent experimental psychologist, Geoffrey R. Loftus, wrote a review article (1996) titled "Psychology Will be a Much Better Science When We Change the Way We Analyze Data," asserting that "Null Hypothesis Statistical Testing, as typically utilized, is barren as a means of transiting from data to conclusions."

Statisticians who are not specialists in psychology might ask: Who are these people, anyway? Where did they get such odd ideas? And why should we care? I think that all three of these questions are worth answering. Psychologists make massive use of statistical inference in research; they tend to perform their own data analyses, to seek help mainly from home-grown experts, and to teach their separate statistics courses to undergraduate majors and to graduate students. This massive branching into specialized teaching and methodology offers interesting lessons about the practice of statistics; may offer novel opportunities for research, teach-

David H. Krantz is Professor of Psychology and Statistics, Columbia University, 618 Mathematics (mail code 4403), New York, NY 10027 (E-mail: *dhk@columbia.edu*).

Journal of the American Statistical Association
December 1999, Vol. 44, No. 448, Review Paper

ing and consulting; and may even suggest weaknesses in the theoretical underpinnings of statistical science.

In the next three sections I develop a more detailed answer to the three questions just framed. To answer the first question, I briefly sketch the diverse views about statistical methods held by different groups of psychologists; this leads to a brief account of the background of the authors of the book under review. For the second question, I outline the reasons that null hypothesis testing is perceived as problematic in psychological research and why confidence intervals are often viewed as a partial solution. The third question is the most important: Statisticians should take notice of this controversy, because it focuses attention on a major gap in the foundations of statistics.

2. PSYCHOLOGISTS' VIEWS OF STATISTICS

Tests of statistical hypotheses (sometimes in the form of confidence intervals) are used as a central methodological element in all, or nearly all, empirical articles in every journal published by the American Psychological Association, the American Psychological Society, or the Psychonomic Society, as well as in most psychology journals published commercially. A clear statement of the rationale for this practice was given by Arthur W. Melton in his valedictory editorial closing out 12 crucial years (1951–1962) as Editor of the *Journal of Experimental Psychology* (which, from its founding in 1916 to the present has been a leading archival journal for empirical research in psychology in the United States). Though his language now sounds old-fashioned, I believe that the statement is still worth reading or rereading today. Melton discussed the standards of "validity, reliability, and substantiality" that he and his associate editors had enforced for acceptance of manuscripts during these 12 years. On the topic of reliability, he wrote:

> The next step in the assessment of an article involved a judgment with respect to the confidence to be placed in the findings—confidence that the results of the experiment would be repeatable under the conditions described. In editing the Journal there has been a strong reluctance to accept and publish results related to the principal concern of the research when those results were significant at the .05 level, whether by a one- or two-tailed test! This has not implied a slavish worship of the .01 level or any other level, as some critics may have implied. Rather, it reflects a belief that it is the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounced. At least, it was believed that such findings do not deserve a place in an archival journal... The P level of a finding which was the major purpose of the investigation... is only one element in the persuasion, others being the relation of necessity between the predicted relationship and other previously or concurrently demonstrated effects, and the consistency of the relationship across a sequence of experiments... The same philosophy applied when negative results were submitted for publication, but here rejection frequently followed the decision that the investigator had not given the data an opportunity to disprove the null hypothesis, i.e., the sensitivity of the experiment was substandard for the type of investigation in question and was therefore not sufficient to persuade an expert in the area that the variable in question did not have an effect as great as other variables of known significant effect (1962, pp. 553–554).

This passage strikes me as remarkable both for its explicit standards and its implicit assumptions. Explicitly, it contains criteria both for reliability of new findings and for reliability of negative findings; that is, for power analysis. Implicitly, it assumes that psychological research is cumulative, so that one generally knows to what effect size is important in a given domain. (The word "significant" in the final sentence is clearly used in the sense of "important" rather than in the commonly encountered sense of "statistically reliable".) Furthermore, it assumes that an *effect* is a change in central tendency of a distribution of measurements, with individual measurements subject to unexplained error. A final implicit assumption is that mathematical models will be used to assess error distributions for estimates of change in central tendency, but the passage does not envisage that the pattern and magnitudes of the effects themselves will be modeled mathematically.

Although this passage still represents one of the principal outlooks on statistical thinking among research psychologists, three other views should be described briefly:

- Nothing is due to chance. This is the Freudian stance (e.g., Freud 1917), and it has at least the merit (similar to exploratory data analysis) of subjecting many small and large occurrences to close scrutiny. Of course, this viewpoint can easily lead to gross overfitting of data. I think that the notion that all of the variance can be explained is implicit in "popular" psychology, but, fortunately, it has little support among researchers.
- Only large effects should be studied. This is another antistatistical stance, emerging from the research tradition of Pavlov and other physiologists. If one chooses a good problem and investigates it with sufficiently clever measurement in a properly chosen setting, then one can magnify signal and shrink noise. The only statistical test one ever needs is the IOTT or "interocular trauma test." The result just hits one between the eyes. If one needs any more statistical analysis, one should be working harder to control sources of error, or perhaps studying something else entirely.
- Develop fully quantitative theories. This stance has given rise to the subfields of psychometrics and mathematical psychology. I cannot take the space to describe the histories of these two branches and their differences. It is sufficient to remark that they share two important features. First, unlike the passage quoted from Melton (1962), they attempt to model patterns and magnitudes of effects, as well as error processes; they share the statistician's view that an "effect" is a parameter value (or change thereof) in a mathematical model. Second, these subfields are a natural source of supply for home-grown experts, who make themselves useful by teaching elementary and advanced statistics and by consulting with their colleagues on problems of data analysis. By no means are all, or even most, of the home-grown statistical experts in psychology specialize in quantitative psychology, but the latter group are by and large the psychologists with the most extensive training in mathematics and statistics.

This account provides the background for the answer to the first question posed earlier. Who are these people? Almost all of the contributors to the book are at least partially identified with the psychometric branch of quantitative psychology, and most have achieved high distinction in psychology. More than half have served at least as consulting editors for the journal *Multivariate Behavioral Research*, which is published under the auspices of the Society of Multivariate Experimental Psychology. This journal publishes theoretical research on multivariate methods, empirical applications of such methods, and methodological papers. To give the flavor, volume 32 (1997) includes (among many other things) a simulation-based examination of the robustness of a multivariate Welch–James test (for unbalanced repeated-measures designs, with between-group heterogeneity of aspherical covariance matrices), a regression-based study of the effects of androgen on genital morphology of neonatal female rats, and a methodological discussion of using means-and-covariance structures analysis to examine cross-cultural constancy of factor structures.

The book discussed here is the first in the planned Multivariate Applications book series, sponsored by this same Society and edited by Lisa L. Harlow. The series "welcomes methodological applications from a variety of disciplines, such as psychology, public health, sociology, education and business."

In short, this group of authors is sophisticated in statistical theory and heavily experienced in both consulting with and teaching psychologists. Many of them also have strong empirical research programs in some area of psychology.

3. THE ABUSE OF HYPOTHESIS TESTING

The overuse and misuse of null hypothesis tests in psychological research has been so thoroughly documented during the past 30 years or so that this volume devotes minimal space to rehearsing these flaws. A brief overview is given in Chapter 2, by Jacob Cohen. This is actually a reprint of Cohen's Lifetime Achievement Award address for the Society of Multivariate Experimental Psychology, titled "The Earth is Round ($p < .05$)." The opening of his abstract sums up the indictment:

> After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of p as the probability that H_0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H_0 one thereby affirms the theory that led to the test.

Here, in two sentences Cohen covers (1) the hypothesis test as ritual, (2) the hypothesis test as substitute for looking at the data, (3) the confusion of $\Pr(D|H)$ with $\Pr(H|D)$, (4) overoptimism about replicability, and (5) confusion of hypothesis rejection with confirmation of a theory. Another error, closely related to (4) and (5), perhaps deserves separate mention: (6) misinterpretation of failure to reject H_0 as failure to replicate an earlier study. Chapter 7, by Joseph S. Rossi, focuses on this latter fallacy.

Many of the authors favor replacing the point hypothesis test by a report of a confidence interval. This simple change deals in part with each of (1)–(6). The extent to which this is so is, curiously, not spelled out in any one place in the volume, although relevant arguments are scattered throughout different chapters. A more systematic account of the benefits of such a change can be found in the book by Oakes (1986).

Briefly, for (2), the confidence interval includes a point estimate and so at least enforces that much of a look at the data (note that some published articles give *only* p values); for (3), the confidence interval has a valid procedural probability interpretation; for (4), a wide confidence interval reveals ignorance and thus helps dampen optimism about replicability; and for (6), if two confidence intervals overlap very heavily, then it is harder to consider one of the experiments as a failure to replicate the other, and, finally, the confidence interval encourages the analyst and the reader to think about the magnitude of the parameter values that are compatible with the current data, which may in turn lead to thinking about the meaning of the parameter and the specificity with which it has been predicted by theory. Thinking is a departure from ritual (1), and specificity of prediction is the key to sharp confirmation of theory (5).

Despairing of change through improved statistical education, some critics want editors of psychology journals to impose an outright test ban. Although it seems difficult to legislate clear thinking, perhaps the argument is that the altered practice will itself contribute to clearer thinking. That, at any rate, has been my own hope—in my statistics classes, for the past 25 years or more, I have advocated the replacement of many hypothesis tests by confidence intervals and have tried to show students what errors can be avoided thereby. I have seen many of these students enter their dissertation research using confidence intervals, but then fall back into poorly interpreted hypothesis tests (because that is all their advisors understand). At times, I too have despaired of jawboning and wished that I could impose a test ban by fiat.

I hope that by now I have answered my second question: Where did (some of) these people get such odd ideas?

4. THEORY AND PRACTICE

Statisticians prove theorems or develop methods that, if properly applied, would be useful. If people misapply them, this is viewed as a problem for education, not for statistical research. The statistician may hasten to put on her or his educator's hat and make a strenuous effort to improve statistical education, but the intrinsic value of statistical methods is judged by their costs and their benefits when properly used, not by the blunders of the poorly educated.

What is missing from this viewpoint is that formal statistical methods are not the whole, but only a part of *inductive inference* (and in many areas of science only a minor part). Some statisticians concerned with visual display of information have in recent years been quite sensitive to the fact that the statistical methods need to take into account the

facts of human perceptual processing (Cleveland 1993). Visual pattern perception is indeed an important part of human inductive inference, but it is far from the whole. In particular, scientists enter their profession as adults, having had many years of practice in causal inference. Among other things, they have discovered what behaviors are likely to produce frowns or smiles from parents, teachers, or lovers; and I think that even statisticians eschew formal calculations in analyzing the data that lead to such practical causal knowledge. Scientific inference often deals with relations more esoteric than those we discover in everyday life and operates in an extended social and institutional setting, but the basic cognitive processes used to invent and confirm scientific hypotheses are probably the same as or are minor elaborations of those used in everyday life.

There has been little or no detailed consideration of how formal statistical methods articulate with the basic cognitive processes of scientific inference. It is not astonishing that statisticians have largely neglected this question; most are trained in mathematics and not in cognitive psychology. Nor could one realistically expect this to be addressed by philosophers of science, who deal mainly with questions far removed from actual scientific process.

The contributors to the present volume, however, have training in psychology as well as in statistics; most are experienced empirical scientists, and many put forward strong claims concerning relations between statistical methodology and scientific inference or scientific progress—claims that would seem to cry out for supporting arguments. One thus might have expected deep insights into the relationship between statistical methodology and science from such a group of authors and in support of the strong claims that they make for particular methods. Yet some of these claims are made without any argument to support them, and in most of the chapters—an exception is Chapter 14, by Paul E. Meehl—the supporting arguments are superficial. I elaborate on this point in Section 6.

This brings me to the third question posed earlier: Why should we care? I answer this by suggesting that neither statisticians nor psychologists (including the authors of this volume) adequately understand the ways in which research psychologists use statistical methods to further their scientific work. It is all too easy for us to denigrate that which we misunderstand, particularly when it does not fit into decision-theoretic conceptions (frequentist or Bayesian) of how scientific inference ought to proceed. Our understanding is also impaired by massive selection bias in the examples that we see as consultants. Based on my own experience, most good research psychologists consult only occasionally with statistical experts. Thus, although experts sometimes see outstanding science in their consultations, they more often see poor practice: an inexperienced scientist, working on a poorly chosen problem, hoping for a statistical miracle. Such situations are quite vivid and far from rare, yet they offer the statistician little insight concerning the effective roles of statistical methods in good scientific work.

A more careful examination of psychological statistics, and especially its successes along with its absurdities, may lead to a better understanding of how statistics functions in real-world inferential settings. This will deepen the foundations of statistics and may lead in turn to some new statistical methods, as well as to better teaching and sounder consultation. Cleveland (1993) has accomplished some of what is needed in his examination of inference from graphs, but it is important to consider more generally how scientists use all sorts of statistical methods in their reasoning. The present volume is valuable to statisticians because it highlights a problem that needs attention, not because it solves that problem.

5. DISTINCT GOALS IN HYPOTHESIS TESTING

As a start toward a deeper investigation of foundations, I offer my own observations concerning the rather distinct uses made of hypothesis testing. From the standpoint of mathematical statistics, these are distinctions without a true difference; this illustrates the power and scope of an abstract framework. From the standpoint of decision-theoretic foundations (whether Bayesian or frequentist) the same is true, but in my view this merely reveals the superficiality of standard decision theory. The confusion among these distinct uses underlies both the abuses of hypothesis testing discussed here and the overreaction of the critics.

5.1 Checking the Adequacy of a Provisional Working Model

This inferential goal is implicit in almost everything people do. We continually seek reassurance that what we believe does indeed remain true. Crossing a street, one checks that automobiles stopped at a red signal remain motionless; speaking, one notes that the listener shows signs of understanding; parting for a day from a lover, one looks for some word or gesture that renews the commitment. In science we have many standard rules or procedures that remind us to check some of our more esoteric beliefs; for example, a vision researcher calibrates light sources frequently, and research psychologists repeat control groups that they have run before and incorporate manipulation checks into their experimental procedures. Apart from such formal checks, scientists show the same alertness in their work as in everyday life to deviations from the expected.

In all of these examples from everyday life and from science, one collects data that are expected to conform to a model and whose conformity to that model would not be worth reporting for its own sake. The vision researcher would not publish daily calibration data in a separate, stand-alone article; the calibrations are reported only incidentally, to validate the results of novel experiments. The control-group data or the manipulation check results would not be published separately either; the only news in these results is that things are as they are expected to be.

Scientists often design experiments in such a way that simple statistical models are expected to be at least approximately valid. For example, psychologists often include

practice trials on a task in part so that the "random" variance is expected to not change very much from earlier trials to later ones. However, the simplifying assumptions incorporated into models must be checked as much as possible, and good scientists devote much informal and formal data analysis to checking them.

Perhaps the most frequent formal tests of this sort in psychological research are tests for equality of variance in two groups, usually done in conjunction with two-sample t tests that may or may not assume such equality. Modern statistical packages provide various facilities for checking the simplifying assumptions of statistical models, and these are used increasingly, though researchers often are not sure what they should do if the simplifying assumptions are rejected.

I note three special features of this sort of hypothesis test that distinguishes it from all of the other types discussed. First, it is often impractical to check some assumptions. The tests are never carried out; rather, conclusions or estimates based on the model must be viewed as conditional on the unverified assumptions. Second, the alternatives to a simplified model may have many additional parameters, so formal tests have low power and only large deviations are apt to be detected. Thus robustness of the model-based inferences is crucial. Third, a continual state of alertness means that in practice a great many implicit tests have taken place. One cannot possibly control simultaneous type I error rate; false alarms in model checking must be regarded with equanimity.

5.2 Evaluating Important Model Parameters

In everyday language, this is called measurement. Physics abounds in examples. Psychological examples include measurement of individual scores on ability or personality scales. A well-known example in cognitive psychology comes from a task in which the subject must decide as rapidly as possible whether or not a presented item (e.g., a digit or a letter) matches any of the items in a previously memorized set (Sternberg 1966). Reaction time increases linearly with the number of items in the memorized set. The parameter to be measured is the slope of this linear function (often around 30–40 msec/item, and sometimes interpreted as the rate of scanning of items in working memory).

The inferential goal is to specify what values are strongly ruled out by the available data, or (in Bayesian approaches) to specify the posterior probability distribution over the possible values, given the available data. The most common statistical calculation is the confidence interval, which is usually equivalent to a series of hypothesis tests with a fixed size of rejection region. Techniques that take many different measurements with different systematic errors into account (e.g., random-effects meta-analysis) may give a more realistic interval.

Such measurements obviously assume the validity of a model. The meaning of the true value of the parameter, and thus of the estimate or measurement, derives from the model.

5.3 Showing That Results That Seem to Confirm a Theory are not Attributable to Mere Chance

This goal accounts for most of the explicit formal use of statistics in psychological research. The researcher formulates a theoretical prediction, generally the direction of an effect (commonly nowadays, the direction of a second-order or higher interaction). When the data in fact show the predicted directional result, this seems to confirm the hypothesis. The researcher tests a "straw-person" null hypothesis that the effect is actually zero. If the latter cannot be rejected at the .05 level (or some variant), then the apparent confirmation of theory cannot be claimed, for the reasons outlined by Melton (1962). Researchers often conclude that a larger sample size or greater control over variability, or both, are needed to produce a clear (and publishable) demonstration. The additional work thus generated is often very informative, one way or another.

In this situation, rejection of the straw-person null hypothesis at the .05 or .01 level should be regarded as only a preliminary step; other methods are needed when it comes to examining the extent to which theory underlying the test has actually been confirmed by the finding. A common error in this type of test is to confuse the significance level actually attained (for rejecting the straw-person null) with the confirmation level attained for the original theory. Statistics could help researchers avoid this error by providing a good alternative measure of degree of confirmation. What is missing (as Meehl points out in his chapter in the book) is a measure of the "riskiness" of theories, (i.e., the sharpness of their numerical predictions), because strength of confirmation actually depends on that, not on the significance level attained for a straw-person null.

A closely related situation arises in much applied research in psychology and other disciplines (e.g., clinical trials). A directional-effect prediction is derived not so much from extending a theory into a new area as from past experience with related research questions. For example, a hypothesis about the effect of a new drug on humans may be based on experience with analogous modifications of drugs and on experience with animal trials. A straw-person null hypothesis test is again just a preliminary step, and in such situations the next steps are often much clearer than in basic research in psychology. For example, one may calculate a confidence interval or posterior credible interval for the magnitude of the effect. The straw-person null hypothesis test becomes an almost useless preliminary. If such straw-person tests are a bit more useful in the context of basic research in psychology, this is because decisive followup steps are more often in doubt and may turn out to be unique to the particular domain being studied.

5.4 Testing a Serious (Approximate) Null Hypothesis

On some occasions, an important scientific theory is cast in the form of an approximate null hypothesis. In the example given earlier from Sternberg's study of the rapid scanning of working memory, linearity of the set-size function (for size ≥ 2) and identity of the set-size slopes for correct detection and correct rejection were tested and con-

firmed (approximately); some of my own research involved rejecting linearity of response of the yellowness-blueness opponent-color response, contrary to then-prevailing theory.

In the previous section I noted that the attained significance level for rejecting a straw-person null hypothesis has little interest, because the main issue is the strength of confirmation of the theory. For the present goal type, however, attained significance level is quite important, if the size of the apparent deviation from the theory is held constant. Rejecting an important theory merely at the .05 level may make ears perk up perhaps, but is usually far from convincing; an attained significance level of .0001, obtained by a much larger sample size and showing the same magnitude of deviation, would be much more convincing and would require that one either reject the theory or reexamine the auxiliary assumptions that underly the relevance of the experiment.

Because such theories are usually considered approximations, rejections involving a sufficiently small effect size are in fact taken as confirmation rather than rejection, no matter what significance level is attained. For example, in my own work mentioned earlier, I could also reject decisively the predictions of linearity for the redness-greenness opponent-color response, but the effect was so small that I in fact drew the opposite conclusion—that the hypothesis was confirmed to a high degree of approximation. In a straw-person null hypothesis test, in contrast, a very small effect size (which nonetheless rejects the null hypothesis because of very large sample size) would ordinarily reject rather than confirm the underlying theory, because the latter predicted a vague (but not very small) directional effect. This kind of occurrence is rare in psychology however; one does not devote that sort of effort to confirming a straw-person null hypothesis.

5.5 Choosing an Appropriate Action or Policy

Metaphorically, all hypothesis testing has been subsumed under this heading, but I believe that it is a serious foundational mistake to pursue this metaphor very far. In none of the four situations discussed earlier would one ever act as though the tested hypothesis is *false* when the attained significance level is .70, nor would one act as though it is *true* when the observed deviation is substantial and the attained significance level is .001 or less. Scientists do not plot 30% confidence intervals, and if 99.9% confidence intervals are ever used, it is just a form of boasting, to show (because even the 99.9% interval is very short) that the measurement is extraordinarily accurate. However, it is easy to find real decision settings in which the costs dictate optimal type I error proportions above .90 or below 10^{-6}.

5.6 Another Look at Criticisms of Null Hypothesis Testing

Let me start with the point made in the preceding section. The rigidity of significance levels in scientific inference has been the target of criticism by statisticians. My view is quite the opposite: it is rather that scientific practice has something important to tell us about the foundations of statistics. We need a foundational theory that encompasses a variety of inferential goals, rather than one that theoretically reduces incommensurable goals to a common currency of "cost" or "utility."

Sections 5.1, 5.3, and 5.4 each illustrate legitimate and valuable uses of null hypothesis testing within psychology. The real trouble is that the straw-person tests (5.3) predominate and they are both misused by investigators, as though the situation in 5.4 obtained (attained significance level is taken seriously), and are misanalyzed foundationally, as though the situations in 5.4 or 5.5 obtained. If such tests were rationalized better, then perhaps they would be used better, particularly if alternative methods were available for assessing strength of confirmation of a theory.

6. STRENGTHS AND WEAKNESSES OF THE BOOK

6.1 Some Good Ideas

Considerable wisdom about statistical methods is scattered about in this volume. The chapters by Cohen, by Abelson, and by Meehl are particularly valuable. Part of Abelson's chapter discusses some of the varied uses of hypothesis tests in scientific inference. He argues that although some tests might well be replaced by confidence intervals, hypothesis tests remain important in model checking. Meehl's chapter title makes his main point: "The Problem is Epistemology, Not Statistics." As mentioned earlier, he suggests the need for a new formal tool, outside the current scope of inferential statistics, to assess the "riskiness" or sharpness of numerical predictions from theories. Two other chapters offer new ideas to help evaluate statistical models: Chapter 8, by Roderick P. MacDonald, discusses the idea that the effect size for deviations from a model can be viewed as a measure of goodness of approximation for that model; Chapter 9, by James H. Steiger and Rachel T. Fouladi, discusses confidence intervals for effect size or for R^2, based on noncentral chi-squared, t, or F distributions. Chapter 13, by William W. Rozeboom, begins with 48 pages of heavy philosophical discussion, but for dessert one gets a delightful 6-page "epilog" with practical advice about applying statistics in research.

6.2 Some Weaker Chapters

I have less favorable assessments of some other chapters. Lisa L. Harlow spends much of Chapter 1 counting the other contributor's noses with respect to various issues. I do not believe that support by m out of n authorities is an argument worth mentioning in favor of any inferential practice.

In Chapter 6, Richard J. Harris uses "three-valued logic" in a sense that is far from the standard usage in formal logic. Charles S. Reichardt and Harry F. Gollub in Chapter 10 offer a bizarrely narrow distinction between Bayesian and frequentist statistical logic: The key idea, according to their presentation, is that frequentists use the word "confidence" wherever Bayesians use "probability."

Robert M. Pruzek's presentation of Bayesian ideas in Chapter 11 discusses only the subjective approach, suggesting that he is out of touch with current Bayesian work; moreover, his discussion includes one awkward mathematical error. In discussing the beta-binomial model, he says:

> Of course, in the event that the prior is tight and the data *contradict* the modal π, the posterior will tend to become more diffuse initially as empirical data are accumulated.

Wishful thinking! Under the beta-binomial model, if the prior mean is in the central interval (1/3, 2/3), the variance of the posterior will always be smaller, no matter what new data are obtained. For example, for prior mean = .35, even if the true value is $\pi = 1$, the variance decreases monotonically as new observations are made. Even for a prior mean outside this interval, the initial diffusion of the posterior occurs only when the truth is quite far from the prior, and even then the effect is limited. For example, let the prior be $\beta(4, 16)$, a rather sharp prior with a mean of .20, and suppose that the truth is $\pi = .80$. At sample size = 10, a perfectly representative (8, 2) sample gives a posterior $\beta(12, 18)$ that is still very far from the truth but already has variance smaller than that of the prior, and additional data can only drive the posterior variance down further.

6.3 Exaggerated Claims

The preceding are relatively minor complaints, however, in comparison with the book's principal defect. Strong factual claims are put forth without any empirical evidence offered in their support (Chapters 3 and 11) or on the basis of evidence that seems to me to have been misinterpreted (Chapter 7). In Chapters 3 and 7, and in other critiques of hypothesis testing, there is an implicit or partially explicit argument that can be analyzed into the following sequence of five assertions:

Step 1: Many or most research psychologists understand statistics poorly and use it in practice in ways that reveal this misunderstanding.

Step 2: Such misunderstandings seriously impair scientific inference.

Step 3: Impaired scientific inference slows scientific progress, or even halts it.

Step 4: Scientific progress in psychology is slow or nonexistent.

Step 5: Therefore, the misunderstanding of statistics is one of the major causes of poor progress in psychology.

Both assertions (1) and (4) are alleged facts, for which documentation might be required, whereas assertions (2), (3), and (5) are causal attributions, which also might be deemed in need of strong supporting arguments.

In my opinion, assertion (1) is sufficiently obvious that detailed documentation may be superfluous. That poor understanding is highly prevalent and that it shows up in the practice of statistics are facts that are as obvious to statistical consultants as the fact that chilly weather is prevalent in much of the United States during the winter season. Because the sample of researchers seen by statistical consultants is quite biased toward misunderstanding and poor practice, it is hard to be sure just how prevalent these misunderstandings are, hence I inserted the hedge "many or most," after which I would endorse this assertion.

But assertion (2) is not a necessary consequence of (1), and it is not an obvious fact of experience either. It is one thing to accuse scientists of showing their ignorance of statistical reasoning in the course of their science, but this does not imply that their ultimate conclusions will be incorrect, nor even that their efficiency in reaching correct conclusions will be impaired. A causal attribution of this sort needs to be supported by careful empirical arguments.

Even if (2) were admitted, assertion (3) is not a necessary consequence of (2). The overall rate of progress could be rapid despite inefficiency or errors in inference. It is at least logically possible that inferential errors could sometimes lead to speedier progress: for example, dissatisfaction with apparently inconsistent results could motivate scientists to devise cleaner observational methods. Here, too, the causal attribution needs careful empirical arguments in support before it is accepted.

I disagree with assertion (4): I believe there has been great progress in many subfields of psychology, and would be astonished if some of the critics of null hypothesis testing have not said the same, particularly in their grant proposals and similar documents. Finally, even if (1)–(4) were all true, one need not draw the causal conclusion (5); other factors could be far more important than statistical misunderstandings in determining the rate of scientific progress. Thus my objection is to making assertions (2)–(5) without presenting empirical arguments for each.

A typical example is a passage from Thompson (1992), quoted by Cohen in Chapter 2:

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired. This tautology has created considerable damage as regards the cumulation of knowledge (p. 436).

I agree with the first sentence: finding $p < .05$ often tells the reader only what the investigator already knows, that great effort was put forward to obtain a large enough sample to compensate for the high noise level and/or modest effect size. I also agree with what may be implied by this first sentence: In most cases, much more could be extracted from the data by thorough analysis. However, I find the leap from the first sentence to the second astonishing. The literature in history and sociology of science suggests that such causal assertions would not be easy to establish, even with hard work to amass relevant evidence.

The opening section of this review quoted a similar example of strong claims, unsupported by evidence or argument, from the abstract to Chapter 3. The chapter itself is a fascinating document. The authors' past works called for a ban on null hypothesis tests. They have now collected a large number of counterarguments offered in response to these calls. The chapter groups these counterarguments un-

der eight main headings, and offers refutations for each in turn.

Although I agree with most of the authors' criticisms of null hypothesis tests and of the counterarguments that they cite, they do misread at least one counterargument, they essentially ignore the counterarguments that Abelson presents in Chapter 5, and they make some extraordinary negative and positive assertions. Their most extreme negative assertion is the flat challenge that significance testing never makes a positive contribution to the research enterprise. Abelson takes up this gauntlet in Chapter 5. In my view, however, such a challenge misunderstands the nature of the contribution of significance testing. It is but one element in a larger process of scientific inference. Scientific inference, like bicycle riding and other skills, can be done very well by people who cannot explain well what it is that they are doing. Theories about scientific inference are still primitive, both in philosophy and psychology, and so it is quite hard to gauge the contribution of one isolated element. I have taught students to replace most significance testing—especially simple tests concerning means—by confidence intervals for contrasts for over 25 years, yet they still continue to look first at whether or not 0 is in the confidence interval! Sometimes I do that too, and when I do, I get a partial, not yet fully articulated insight into the actual function of significance testing in scientific thought. In physics research, a theoretical nonzero value is often compared with a confidence interval. The authors claim that this is not significance testing, but I think that everyone looks immediately to see whether or not the interval includes the theoretical value. That to me is a significance test, albeit one with a very different goal from most of those in psychology.

The strongest positive claim in Chapter 3 is for the benefits of meta-analysis:

> Only by combining findings across multiple studies using meta-analysis can dependable scientific conclusions be reached (p. 52).

Such an assertion demands empirical support. The authors' past publications (Hunter and Schmidt 1990; Schmidt 1996), though valuable in many respects, are unconvincing to me in this regard. No examples are cited in Chapter 3. I discuss an example in Chapter 7 later.

A final empirical claim, which I find quite disturbing, is that those who disagree with them are motivated by factors other than a desire to know the truth:

> Accepting the proposition that significance testing should be discontinued... entails the difficult effort of changing the beliefs and practices of a lifetime (p. 49).
> The main lesson specific to this chapter is this: Beware of plausible and intuitively appealing objections to discontinuing the use of significance testing... Finally, try to show enough intellectual courage and honesty to reject the use of significance tests despite the pressures of social convention to the contrary. Scientists must have integrity (p. 61).

The adoption of new techniques or concepts has always been effortful for experienced researchers; yet in psychology, as in other sciences, large numbers have repeatedly made such efforts and have changed the "beliefs and practices of a lifetime." Thus there is plenty of evidence that failures of courage and honesty are *not* the principal factors in this continuing saga, and the authors offer no evidence to support the claim that lack of integrity is a major cause of psychologists' adherence to significance testing.

On the contrary, I take this causal claim as just another symptom of the fact that the authors (in common with everyone else, certainly including me) simply do not yet very thoroughly understand the cognitive processes involved in scientific inference. When these processes are better understood, we may be better able to explain (and combat) scientists' excessive reliance on significance testing.

The flaws in Chapter 3 are really a pity, because the authors' analysis of many weak defenses of null hypothesis testing could be very valuable in helping some who are confused to understand the issues.

A rather different sort of unsupported empirical claim is Robert Pruzek's assertion in Chapter 11 about the efficacy of subjective Bayesian methods:

> More often than not, experience has shown that when investigators are honest in forming their priors, taking all relevant information into account at the outset, strong contradictions between subjectively based priors and data-based likelihoods tend to be rare (p. 298).

The author cites no data, and I know of no empirical studies that directly probe the accuracy of subjective priors in scientific investigation. Contrariwise, there is a substantial empirical literature showing overconfidence (poor coverage properties) for priors obtained in other domains (see, e.g., Lichtenstein and Fischhoff 1980 for general findings, Christensen-Szalanski and Bushyhead 1981 for overconfidence in physicians' judgments, and Yaniv and Foster 1997 for an explanatory theory). Results of Li and Krantz (1996) suggest that overconfidence also would hold for priors in a social-science domain. In any case, this sort of statement should be a red flag for the skeptical reader, who might wonder how easily pertinent studies could be designed. How should one determine whether an investigator has been "honest" in forming a subjective prior and whether "all relevant information" was taken into account in forming the prior?

Chapter 7, "A Case Study in the Failure of Psychology as a Cumulative Science: The Spontaneous Recovery of Verbal Learning," by Joseph S. Rossi, makes claims similar to those of Chapter 3, but the arguments for those claims are either omitted or seriously flawed. The chapter claims that psychology is failing as a cumulative science and that reliance on dichotomous interpretations of significance levels, rather than more apt statistical methods, has contributed substantially to this failure. A major section illustrates the nefarious influence of dichotomous interpretation by showing how the subfield of human memory concluded that a particular phenomenon—spontaneous recovery of old associations—is unreliable or ephemeral. Rossi's meta-analysis suggests an effect size of .27–.48 (95% confidence); the number of nonsignificant results in the literature (more than half) could have been expected, given the power of those studies for an effect size of .39.

For the claim of failure as a cumulative science, Rossi does not cite any concrete evidence, merely the opinion of

Meehl (1978). This work by Meehl is in fact highly relevant to the present discussion. It restricts its claim of cumulative failure to "soft" psychology (which has never been taken to include the field of human memory). Even within that domain, Meehl does not really document failure; he is anecdotal rather than systematic, and suggests five areas of lasting contribution within soft psychology as it stood at that time. Moreover, Meehl's article offers a list of 20 distinct factors that compete with poor statistical methodology as explanations of slow progress in psychology. Meehl clearly explains the flaws in the use of null hypothesis tests, but though his title seems to imply that this flawed practice is more important in its scientific consequences than many of his 20 alternative explanatory factors, he gives no actual arguments for that opinion.

As for the particular illustration, involving the phenomenon of spontaneous recovery of old associations, I am afraid that Rossi's arguments illustrate the dangers of poor meta-analysis much more clearly than the dangers of significance testing.

Human memory is one of the most highly developed subfields of psychology, in terms of empirical facts and sophisticated theories. Spontaneous recovery of associations was not studied primarily as a phenomenon in its own right; rather, it was predicted by a theory (unlearning theory) that was seriously considered for a while in connection with particular facts about forgetting. The prediction was not merely that spontaneous recovery would occur; rather, it included specifics about its detailed temporal course. Attempts to verify this prediction failed: Spontaneous recovery did not occur at the predicted times, but occurred at times much too short to be counted as support for the theory. Subsequently, other findings thoroughly discredited unlearning theory, and this undercut much of the motivation to continue investigation of what was now a small and theoretically isolated effect.

Note that the difference from the theoretically expected time course not only reduced the theoretical interest in the effect, but also might explain some of the failures to find the effect. Rossi's chapter does not touch on this critical issue, however. Lumping together findings at different time intervals, without attention to their theoretical significance, simply does not constitute an appropriate meta-analysis in this literature.

I now summarize the status of Rossi's argument in terms of the five-assertion argument outlined at the start of this section. The chapter does not actually offer any direct evidence for assertion (1), that these scientists did misunderstand hypothesis testing, or for assertion (2), that such misunderstanding led to impaired inference. Although I am willing to accept that misunderstanding is widespread, it does need to be established in any particular case where it is to be held responsible. I am somewhat skeptical that this particular group of scientists did exhibit serious misunderstanding of hypothesis testing.

The evidence adduced does, on the face of it, attempt to support the latter three claims: that scientific inference was in fact impaired across this series of studies, that this impaired inference slowed scientific progress in the subfield of verbal learning, and that the subfield has in fact made poor scientific progress. I would suggest, however, that a deeper understanding of theoretical and empirical developments in this subarea should lead to dismissal of all three claims. In addition, no attempt was made to rule out alternative explanations for the supposed slow progress of this subfield.

7. CONCLUSIONS

The most valuable part of the book here reviewed is its title. For teachers of statistics, it offers some shock value. Teachers who place the logic of null hypothesis significance testing more or less on a par with scientific logic need to be awakened quite rudely; others can at least use the title to make students sit up and listen. If one is looking for a book from which to assign readings, however, the critique by Oakes (1986) might be preferred. The present volume is livelier but also less systematic, less didactic, and less reliable than that earlier work. (One must overlook Oakes' highly critical coverage of meta-analysis, which was possibly a fair representation of its use in psychology 15 years ago and could serve to counterbalance the exaggerated positive views of Hunter and Schmidt, but which failed to anticipate the further development of this method.)

For statisticians, cognitive psychologists, and philosophers, the title states a question whose answer requires a much deeper analysis of scientific inference than has been given heretofore. Significance tests, explicit or implicit, formal or informal, play many different roles in scientific inference, including some for which they are poorly fitted. These varied uses and roles are hinted at in the chapters by Abelson and by Meehl, but nowhere analyzed in depth. My own tentative list of distinct roles for hypothesis tests is given in Section 5. I am unsure whether a unified foundational treatment encompassing these varied uses is possible. However, I believe that clear rationales for hypothesis testing (unified or not) should replace murky decision-theoretic metaphors, and that this replacement will facilitate improvements in both teaching and practice.

[Received March 1999. Revised May 1999.]

REFERENCES

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press.

Christensen-Szalanski, J., and Bushyhead, J. B. (1981), "Physicians' Use of Probabilistic Information in a Real Clinical Setting," *Journal of Experimental Psychology: Human Perception and Performance*, 7, 928–935.

Freud, S. (1917), *Psychopathology of Everyday Life*, New York: Macmillan.

Harlow, Lisa L., Mulaik, Stanley A., and Steiger, James H. (eds.) (1997), *What If There Were No Significance Tests?*, Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J. E., and Schmidt, F. L. (1990), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Newbury Park, CA: Sage.

Li, Y., and Krantz, D. H. (1996), "Overconfidence and the Goals of Interval Estimation," *Journal of Mathematical Psychology*, 40, 362–363.

Lichtenstein, S., and Fischhoff, B. (1980), "Training for Calibration," *Organizational Behavior and Human Performance*, 26, 149–171.

Loftus, G. R. (1996), "Psychology Will be a Much Better Science When We Change the Way We Analyze Data," *Current Directions in Psychology*, 5, 161–171.

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.

Melton, A. W. (1962), "Editorial," *Journal of Experimental Psychology*, 64, 553–557.

Oakes, M. W. (1986), *Statistical Inference: A Commentary for the Social and Behavioural Sciences*, New York: Wiley.

Rozeboom, W. W. (1960), "The Fallacy of the Null Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428.

Schmidt, F. L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers," *Psychological Methods*, 1, 115–129.

Sternberg, S. (1966), "High-Speed Scanning in Human Memory," *Science*, 153, 652–654.

Thompson, B. (1992), "Two and One-Half Decades of Leadership in Measurement and Evaluation," *Journal of Counseling and Development*, 70, 434–438.

Yaniv, I., and Foster, D. P. (1997), "Precision and Accuracy of Judgmental Estimation," *Journal of Behavioral Decision Making*, 10, 21–32.