Psychological Bulletin
1960, Vol. 57, No. 5, 416–428

THE FALLACY OF THE NULL-HYPOTHESIS SIGNIFICANCE TEST

WILLIAM W. ROZEBOOM
St. Olaf College

The theory of probability and statistical inference is various things to various people. To the mathematician, it is an intricate formal calculus, to be explored and developed with little professional concern for any empirical significance that might attach to the terms and propositions involved. To the philosopher, it is an embarrassing mystery whose justification and conceptual clarification have remained stubbornly refractory to philosophical insight. (A famous philosophical epigram has it that induction [a special case of statistical inference] is the glory of science and the scandal of philosophy.) To the experimental scientist, however, statistical inference is a research instrument, a processing device by which unwieldy masses of raw data may be refined into a product more suitable for assimilation into the corpus of science, and in this lies both strength and weakness. It is strength in that, as an ultimate *consumer* of statistical methods, the experimentalist is in position to demand that the techniques made available to him conform to his actual needs. But it is also weakness in that, in his need for the tools constructed by a highly technical formal discipline, the experimentalist, who has specialized along other lines, seldom feels competent to extend criticisms or even comments; he is much more likely to make unquestioning application of procedures learned more or less by rote from persons assumed to be more knowledgeable of statistics than he. There is, of course, nothing surprising or reprehensible about this—one need not understand the principles of a complicated tool in order to make effective use of it, and the research scientist can no more be expected to have sophistication in the theory of statistical inference than he can be held responsible for the principles of the computers, signal generators, timers, and other complex modern instruments to which he may have recourse during an experiment. Nonetheless, this leaves him particularly vulnerable to misinterpretation of his aims by those who build his instruments, not to mention the ever present dangers of selecting an inappropriate or outmoded tool for the job at hand, misusing the proper tool, or improvising a tool of unknown adequacy to meet a problem not conforming to the simple theoretical situations in terms of which existent instruments have been analyzed. Further, since behaviors once exercised tend to crystallize into habits and eventually traditions, it should come as no surprise to find that the tribal rituals for data-processing passed along in graduate courses in experimental method should contain elements justified more by custom than by reason.

In this paper, I wish to examine a dogma of inferential procedure which, for psychologists at least, has attained the status of a religious conviction. The dogma to be scrutinized is the "null-hypothesis significance test" orthodoxy that passing statistical judgment on a scientific hypothesis by means of experimental observa-

tion is a decision procedure wherein one rejects or accepts a null hypothesis according to whether or not the value of a sample statistic yielded by an experiment falls within a certain predetermined "rejection region" of its possible values. The thesis to be advanced is that despite the awesome pre-eminence this method has attained in our experimental journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research. This is not a particularly original view—traditional null-hypothesis procedure has already been superceded in modern statistical theory by a variety of more satisfactory inferential techniques. But the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene.

To examine the method in question in greater detail, and expose some of the discomfitures to which it gives rise, let us begin with a hypothetical case study.

A Case Study in Null-Hypothesis Procedure; or, A Quorum of Embarrassments

Suppose that according to the theory of behavior, T_0, held by most right-minded, respectable behaviorists, the extent to which a certain behavioral manipulation M facilitates learning in a certain complex learning situation C should be null. That is, if "ϕ" designates the degree to which manipulation M facilitates the acquisition of habit H under circumstances C, it follows from the orthodox theory T_0 that $\phi = 0$. Also suppose, however, that a few radicals have persistently advocated an alternative theory T_1 which entails, among other things, that the facilitation of H by M in circumstances C should be appreciably greater than zero, the precise extent being dependent upon the values of certain parameters in C. Finally, suppose that Igor Hopewell, graduate student in psychology, has staked his dissertation hopes on an experimental test of T_0 against T_1 on the basis of their differential predictions about the value of ϕ.

Now, if Hopewell is to carry out his assessment of the comparative merits of T_0 and T_1 in this way, there is nothing for him to do but submit a number of *S*s to manipulation M under circumstances C and compare their efficiency at acquiring habit H with that of comparable *S*s who, under circumstances C, have *not* been exposed to manipulation M. The difference, d, between experimental and control *S*s in average learning efficiency may then be taken as an operational measure of the degree, ϕ, to which M influences acquisition of H in circumstances C. Unfortunately, however, as any experienced researcher knows to his sorrow, the interpretation of such an observed statistic is not quite so simple as that. For the observed dependent variable d, which is actually a performance measure, is a function not only of the extent to which M influences acquisition of H, but of many additional major and minor factors as well. Some of these, such as deprivations, species, age, laboratory conditions, etc., can be removed from consideration by holding them essentially constant. Others, however, are not so easily controlled, especially those customarily subsumed under the headings of "individual differences" and "errors of measurement." To

curtail a long mathematical story, it turns out that with suitable (possibly justified) assumptions about the distributions of values for these uncontrolled variables, the manner in which they influence the dependent variable, and the way in which experimental and control *S*s were selected and manipulated, the observed sample statistic d may be regarded as the value of a normally distributed random variate whose average value is ϕ and whose variance, which is independent of ϕ, is unbiasedly estimated by the square of another sample statistic, s, computed from the data of the experiment.[1]

The import of these statistical considerations for Hopewell's dissertation, of course, is that he will not be permitted to reason in any simple way from the observed d to a conclusion about the comparative merits of of T_0 and T_1. To conclude that T_0, rather than T_1, is correct, he must argue that $\phi=0$, rather than $\phi>0$. But the observed d, whatever its value, is logically compatible both with the hypothesis that $\phi=0$ and the hypothesis that $\phi>0$. How then, can Hopewell use his data to make a comparison of T_0 and T_1? As a well-trained student, what he *does*, of course, is to divide d by s to obtain what, under H_0, is a t statistic, consult a table of the t distributions under the appropriate degrees-of-freedom, and announce his experiment as disconfirming or supporting T_0, respectively, according to whether or not the discrepancy between d and the zero value expected under T_0 is "statistically significant"—i.e., whether or not the observed value of d/s falls outside of the interval between two extreme percentiles (usually the 2.5th and 97.5th) of the t distribution with that df. If asked by his dissertation committee to justify this behavior, Hopewell would rationalize something like the following (the more honest reply, that this is what he has been taught to do, not being considered appropriate to such occasions):

> In deciding whether or not T_0 is correct, I can make two types of mistakes: I can reject T_0 when it is in fact correct [Type I error], or I can accept T_0 when in fact it is false [Type II error]. As a scientist, I have a professional obligation to be cautious, but a 5% chance of error is not unduly risky. Now if all my statistical background assumptions are correct, then, if it is really true that $\phi=0$ as T_0 says, there is only one chance in 20 that my observed statistic d/s will be smaller than $t_{.025}$ or larger than $t_{.975}$, where by the latter I mean, respectively, the 2.5th and 97.5th percentiles of the t distribution with the same degrees-of-freedom as in my experiment. Therefore, if I reject T_0 when d/s is smaller than $t_{.025}$ or larger than $t_{.975}$, and accept T_0 otherwise, there is only a 5% chance that I will reject T_0 incorrectly.

If asked about his Type II error, and why he did not choose some other rejection region, say between $t_{.475}$ and $t_{.525}$, which would yield the same probability of Type I error, Hopewell should reply that although he has no way to compute his probability of Type II error under the assumptions traditionally authorized by null-hypothesis procedure, it is presumably minimized by taking the rejection region at the extremes of the t distribution.

Let us suppose that for Hopewell's data, $d=8.50$, $s=5.00$, and $df=20$. Then $t_{.975}=2.09$ and the acceptance region for the null hypothesis $\phi=0$ is $-2.09<d/s<2.09$, or $-10.45<d<10.45$. Since d does fall within this region, standard null-hypothesis decision procedure, which I shall henceforth abbreviate "NHD," dictates that the experiment is to be reported

[1] s is here the estimate of the standard error of the difference in means, not the estimate of the individual *SD*.

as supporting theory T_0. (Although many persons would like to conceive NHD testing to authorize only rejection of the hypothesis, not, in addition, its acceptance when the test statistic fails to fall in the rejection region, if failure to reject were not taken as grounds for acceptance, then NHD procedure would involve no Type II error, and no justification would be given for taking the rejection region at the extremes of the distribution, rather than in its middle.) But even as Hopewell reaffirms T_0 in his dissertation, he begins to feel uneasy. In fact, several disquieting thoughts occur to him:

1. Although his test statistic falls within the orthodox acceptance region, a value this divergent from the expected zero should nonetheless be encountered less than once in 10. To argue in favor of a hypothesis on the basis of data ascribed a p value no greater than .10 (i e., 10%) by that hypothesis certainly does not seem to be one of the more impressive displays of scientific caution.

2. After some belated reflection on the details of theory T_1, Hopewell observes that T_1 not only predicts that $\phi > 0$, but with a few simplifying assumptions no more questionable than is par for this sort of course, the value that ϕ should have can actually be computed. Suppose the value derived from T_1 in this way is $\phi = 10.0$. Then, rather than taking $\phi = 0$ as the null hypothesis, one might just as well take $\phi = 10.0$; for under the latter, $(d - 10.0)/s$ is a 20 df t statistic, giving a two-tailed, 95% significance, acceptance region for $(d - 10.0)/s$ between $-.209$ and 2.09. That is, if one lets T_1 provide the null hypothesis, it is accepted or rejected according to whether or not $-.45 < d < 20.45$, and by this latter test, therefore, Hopewell's data must be taken to support T_1—in fact, the likelihood under T_1 of obtaining a test statistic this divergent from the expected 10.0 is a most satisfactory three chances in four. Thus it occurs to Hopewell that had he chosen to cast his professional lot with the T_1-ists by selecting $\phi = 10.0$ as his null hypothesis, he could have made a strong argument in favor of T_1 by precisely the same line of statistical reasoning he has used to support T_0 under $\phi = 0$ as the null hypothesis. That is, he could have made an argument that persons partial to T_1 would regard as strong. For behaviorists who are already convinced that T_0 is correct would howl that since T_0 is the dominant theory, only $\phi = 0$ is a legitimate null hypothesis. (And is it not strange that what constitutes a valid statistical argument should be dependent upon the majority opinion about behavior theory?)

3. According to the NHD test of a hypothesis, only two possible final outcomes of the experiment are recognized—either the hypothesis is rejected or it is accepted. In Hopewell's experiment, all possible values of d/s between -2.09 and 2.09 have the same interpretive significance, namely, indicating that $\phi = 0$, while conversely, all possible values of d/s greater than 2.09 are equally taken to signify that $\phi \neq 0$. But Hopewell finds this disturbing, for of the various possible values that d/s might have had, the significance of $d/s = 1.70$ for the comparative merits of T_0 and T_1 should surely be more similar to that of, say, $d/s = 2.10$ than to that of, say, $d/s = -1.70$.

4. In somewhat similar vein, it also occurs to Hopewell that had he opted for a somewhat riskier confidence level, say a Type I error of 10% rather than 5%, d/s would have fallen outside the region of accept-

ance and T_0 would have been rejected. Now surely the degree to which a datum corroborates or impugns a proposition should be independent of the datum-assessor's personal temerity. Yet according to orthodox significance-test procedure, whether or not a given experimental outcome supports or disconfirms the hypothesis in question depends crucially upon the assessor's tolerance for Type I risk.

Despite his inexperience, Igor Hopewell is a sound experimentalist at heart, and the more he reflects on these statistics, the more dissatisfied with his conclusions he becomes. So while the exigencies of graduate circumstances and publication requirements urge that his dissertation be written as a confirmation of T_0, he nonetheless resolves to keep an open mind on the issue, even carrying out further research if opportunity permits. And reading his experimental report, so of course would we—has any responsible scientist ever made up his mind about such a matter on the basis of a single experiment? Yet in this obvious way we reveal how little our actual inferential behavior corresponds to the statistical procedure to which we pay lip-service. For if we did, in fact, accept or reject the null hypothesis according to whether the sample statistic falls in the acceptance or in the rejection region, then there would be no replications of experimental designs, no multiplicity of experimental approaches to an important hypothesis—a single experiment would, by definition of the method, make up our mind about the hypothesis in question. And the fact that in actual practice, a single finding seldom even tempts us to such closure of judgment reveals how little the conventional model of hypothesis testing fits our actual evaluative behavior.

Decisions vs. Degrees of Belief

By now, it should be obvious that something is radically amiss with the traditional NHD assessment of an experiment's theoretical import. Actually, one does not have to look far in order to find the trouble—it is simply a basic misconception about the purpose of a scientific experiment. The null-hypothesis significance test treats acceptance or rejection of a hypothesis as though these were *decisions* one makes on the basis of the experimental data—i.e., that we elect to adopt one belief, rather than another, as a result of an experimental outcome. *But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested.* And even if the purpose of the experiment *were* to reach a decision, it could not be a decision to accept or reject the hypothesis, for decisions are voluntary commitments to action—i.e., are *motor* sets—whereas acceptance or rejection of a hypothesis is a *cognitive* state which may provide the basis for rational decisions, but is not itself arrived at by such a decision (except perhaps indirectly in that a decision may initiate further experiences which influence the belief).

The situation, in other words, is as follows: As scientists, it is our professional obligation to reason from available data to explanations and generalities—i.e., beliefs—which are supported by these data. But belief in (i.e., acceptance of) a proposition is not an all-or-none affair; rather, it is a matter of degree, and the extent to which a person believes or accepts a

proposition translates pragmatically into the extent to which he is willing to commit himself to the behavioral adjustments prescribed for him by the meaning of that proposition. For example, if that inveterate gambler, Unfortunate Q. Smith, has complete confidence that War Biscuit will win the fifth race at Belmont, he will be willing to accept any odds to place a bet on War Biscuit to win; for if he is absolutely *certain* that War Biscuit will win, then odds are irrelevant—it is simply a matter of arranging to collect some winnings after the race. On the other hand, the more that Smith has doubts about War Biscuit's prospects, the higher the odds he will demand before betting. That is, the *extent* to which Smith accepts or rejects the hypothesis that War Biscuit will win the fifth at Belmont is an important determinant of his betting decisions for that race.

Now, although a scientist's data supply *evidence* for the conclusions he draws from them, only in the unlikely case where the conclusions are logically deducible from or logically incompatible with the data do the data warrant that the conclusions be entirely accepted or rejected. Thus, e.g., the fact that War Biscuit has won all 16 of his previous starts is strong evidence in favor of his winning the fifth at Belmont, but by no means warrants the unreserved acceptance of this hypothesis. More generally, the data available confer upon the conclusions a certain *appropriate degree of belief*, and it is the inferential task of the scientist to pass from the data of his experiment to whatever *extent* of belief these and other available information justify in the hypothesis under investigation. In particular, the proper inferential procedure is *not* (except in the deductive case) a matter of deciding to accept (without qualification) or reject (without qualification) the hypothesis: even if adoption of a belief were a matter of voluntary action—which it is not—neither such extremes of belief or disbelief are appropriate to the data at hand. As an example of the disastrous consequences of an inferential procedure which yields only two judgment values, acceptance and rejection, consider how sad the plight of Smith would be if, whenever weighing the prospects for a given race, he always worked himself into either supreme confidence or utter disbelief that a certain horse will win. Smith would rapidly impoverish himself by accepting excessively low odds on horses he is certain will win, and failing to accept highly favorable odds on horses he is sure will lose. In fact, Smith's two judgment values need not be *extreme* acceptance and rejection in order for his inferential procedure to be maladaptive. All that is required is that the degree of belief arrived at be in general inappropriate to the likelihood conferred on the hypothesis by the data.

Now, the notion of "degree of belief appropriate to the data at hand" has an unpleasantly vague, subjective feel about it which makes it unpalatable for inclusion in a formalized theory of inference. Fortunately, a little reflection about this phrase reveals it to be intimately connected with another concept relating conclusion to evidence which, though likewise in serious need of conceptual clarification, has the virtues both of intellectual respectability and statistical familiarity. I refer, of course, to the *likelihood*, or *probability*, conferred upon a hypothesis by available evidence. Why should not Smith *feel*

certain, in view of the data available, that War Biscuit will win the fifth at Belmont? Because it *is* not certain that War Biscuit will win. More generally, what determines how strongly we should accept or reject a proposition is the probability given to this hypothesis by the information at hand. For while our voluntary actions (i.e., decisions) are determined by our intensities of belief in the relevant propositions, not by their actual probabilities, expected utility is maximized when the cognitive weights given to potential but not yet known-for-certain pay-off events are represented in the decision procedure by the probabilities of these events. We may thus relinquish the concept of "appropriate degree of belief" in favor of "probability of the hypothesis," and our earlier contention about the nature of data-processing may be rephrased to say that the proper inferential task of the experimental scientist is not a simple acceptance or rejection of the tested hypothesis, but determination of the probability conferred upon it by the experimental outcome. This likelihood of the hypothesis relative to whatever data are available at the moment will be an important determinant for decisions which must currently be made, but is not itself such a decision and is entirely subject to revision in the light of additional information.

In brief, what is being argued is that the scientist, whose task is not to prescribe actions but to establish rational beliefs upon which to base them, is fundamentally and inescapably committed to an explicit concern with the problem of inverse probability. What he wants to know is how plausible are his hypotheses, and he is interested in the probability ascribed by a hypothesis to an observed experimental outcome only to the extent he is able to reason backwards to the likelihood of the hypothesis, given this outcome. Put crudely, no matter how improbable an observation may be under the hypothesis (and when there are an infinite number of possible outcomes, the probability of any particular one of these is, usually, infinitely small—the familiar p value for an observed statistic under a hypothesis H is not actually the probability of that outcome under H, but a partial integral of the probability-density function of possible outcomes under H), it is still confirmatory (or at least nondisconfirmatory, if one argues from the data to rejection of the background assumptions) so long as the likelihood of the observation is even smaller under the alternative hypotheses. To be sure, the theory of hypothesis-likelihood and inverse probability is as yet far from the level of development at which it can furnish the research scientist with inferential tools he can apply mechanically to obtain a definite likelihood estimate. But to the extent a statistical method does not at least move in the *direction* of computing the probability of the hypothesis, given the observation, that method is not truly a method of *inference*, and is unsuited for the scientist's cognitive ends.

The Methodological Status of the Null-Hypothesis Significance Test

The preceding arguments have, in one form or another, raised several doubts about the appropriateness of conventional significance-test decision procedure for the aims it is supposed to achieve. It is now time to bring these charges together in an explicit bill of indictment.

1. The null-hypothesis significance

test treats "acceptance" or "rejection" of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a *degree* of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true.

2. It might be argued that the NHD test may nonetheless be regarded as a legitimate decision procedure if we translate "acceptance (rejection) of the hypothesis" as meaning "acting as though the hypothesis were true (false)." And to be sure, there are many occasions on which one must base a course of action on the credibility of a scientific hypothesis. (Should these data be published? Should I devote my research resources to and become identified professionally with this theory? Can we test this new Z bomb without exterminating all life on earth?) But such a move to salvage the traditional procedure only raises two further objections. (*a*) While the scientist—i.e., the person—must indeed make decisions, his *science* is a systematized body of (probable) *knowledge*, not an accumulation of decisions. The end product of a scientific investigation is a degree of confidence in some set of propositions, which then constitutes a *basis* for decisions. (*b*) Decision theory shows the NHD test to be woefully inadequate as a decision procedure. In order to decide most effectively when or when not to act as though a hypothesis is correct, one must know both the probability of the hypothesis under the data available and the utilities of the various decision outcomes (i.e., the values of accepting the hypothesis when it is true, of accepting it when it is false, of rejecting it when it is true, and of rejecting it when it is false). But traditional NHD procedure pays no attention to utilities at all, and considers the probability of the hypothesis, given the data—i.e., the inverse probability—only in the most rudimentary way (by taking the rejection region at the extremes of the distribution rather than in its middle). Failure of the traditional significance test to deal with inverse probabilities invalidates it not only as a method of rational inference, but *also* as a useful decision procedure.

3. The traditional NHD test unrealistically limits the significance of an experimental outcome to a mere two alternatives, confirmation or disconfirmation of the null hypothesis. Moreover, the transition from confirmation to disconfirmation as a function of the data is discontinuous —an arbitrarily small difference in the value of the test statistic can change its significance from confirmatory to disconfirmatory. Finally, the point at which this transition occurs is entirely gratuitous. There is absolutely no reason (at least provided by the method) why the point of statistical "significance" should be set at the 95% level, rather than, say the 94% or 96% level. Nor does the fact that we sometimes select a 99% level of significance, rather than the usual 95% level, mitigate this objection—one is as arbitrary as the other.

4. The null-hypothesis significance test introduces a strong bias in favor of one out of what may be a large number of reasonable alternatives. When sampling a distribution of unknown mean μ, different assumptions about the value of μ furnish an infi-

nite number of alternate null hypotheses by which we might assess the sample mean, and whichever hypothesis is selected is thereby given an enormous, in some cases almost insurmountable, advantage over its competitors. That is, NHD procedure involves an inferential double standard—the favored hypothesis is held innocent unless proved guilty, while any alternative is held guilty until no choice remains but to judge it innocent. What is objectionable here is not that some hypotheses are held more resistant to experimental extinction than others, but that the differential weighing is an all-or-none side effect of a personal choice, and especially, that the method *necessitates* one hypothesis being favored over all the others. In the classical theory of inverse probability, on the other hand, all hypotheses are treated on a par, each receiving a weight (i.e., its "a priori" probability) which reflects the credibility of that hypothesis on grounds other than the data being assessed.

5. Finally, if anything can reveal the practical irrelevance of the conventional significance test, it should be its failure to see genuine application to the inferential behavior of the research scientist. Who has ever given up a hypothesis just because one experiment yielded a test statistic in the rejection region? And what scientist in his right mind would ever feel there to be an appreciable difference between the interpretive significance of data, say, for which one-tailed $p = .04$ and that of data for which $p = .06$, even though the point of "significance" has been set at $p = .05$? In fact, the reader may well feel undisturbed by the charges raised here against traditional NHD procedure precisely because, without perhaps realizing it, he has never taken the method seriously anyway. Paradoxically, it is often the most firmly institutionalized tenet of faith that is most susceptible to untroubled disregard—in our culture, one must early learn to live with sacrosanct verbal formulas whose import for practical behavior is seldom heeded. I suspect that the primary reasons why null-hypothesis significance testing has attained its current ritualistic status are (*a*) the surcease of methodological insecurity afforded by having an inferential algorithm on the books, and (*b*) the fact that a by-product of the algorithm is so useful, and its end product so obviously inappropriate, that the latter can be ignored without even noticing that this has, in fact, been done. What has given the traditional method its spurious feel of usefulness is that the *first*, and by far most laborious, step in the procedure, namely, estimating the probability of the experimental outcome under the assumption that a certain hypothesis is correct, is also a crucial first step toward what one is genuinely concerned with, namely, an idea of the likelihood of that hypothesis, given this experimental outcome. Having obtained this most valuable statistical information under pretext of carrying through a conventional significance test, it is then tempting, though of course quite inappropriate, to heap honor and gratitude upon the method while overlooking that its actual *result*, namely, a decision to accept or reject, is not used at all.

Toward a More Realistic Appraisal of Experimental Data

So far, my arguments have tended to be aggressively critical—one can hardly avoid polemics when butchering sacred cows. But my purpose is

not just to be contentious, but to help clear the way for more realistic techniques of data assessment, and the time has now arrived for some constructive suggestions. Little of what follows pretends to any originality; I merely urge that ongoing developments along these lines should receive maximal encouragement.

For the statistical theoretician, the following problems would seem to be eminently worthy of research:

1. Of supreme importance for the theory of probability is analysis of what we mean by a proposition's "probability," relative to the evidence provided. Most serious students of the philosophical foundations of probability and statistics agree (cf. Braithwaite, pp. 119f.) that the probability of a proposition (e.g., the probability that the General Theory of Relativity is correct) does not, prima facie, seem to be the same sort of thing as the probability of an event-class (e.g., the probability of getting a head when this coin is tossed). Do the statistical concepts and formulas which have been developed for probabilities of the latter kind also apply to hypothesis likelihoods? In particular, are the probabilities of hypotheses quantifiable at all, and for the theory of inverse probability, do Bayes' theorem and its probability-density refinements apply to hypothesis probabilities? These and similar questions are urgently in need of clarification.

2. If we are willing to assume that Bayes' theorem, or something like it, holds for hypothesis probabilities, there is much that can be done to develop the classical theory of inverse probability. While computation of inverse probabilities turns essentially upon the parametric a priori probability function, which states the probability of each alternative hypothesis in the set under consideration prior to the outcome of the experiment, it should be possible to develop theorems which are invariant over important subclasses of a priori probability functions. In particular, the difference between the a priori probability function and the "a posteriori" probability function (i.e., the probabilities of the alternative hypotheses after the experiment), perhaps analyzed as a difference in "information," should be a potentially fruitful source of concepts with which to explore such matters as the "power" or "efficiency" of various statistics, the acquisition of inductive knowledge through repeated experimentation, etc. Another problem which seems to me to have considerable import, though not one about which I am sanguine, is whether inverse-probability theory can significantly be extended to hypothesis-probabilities, given knowledge which is only probabilistic. That is, can a theory of sentences of form "The probability of hypothesis H, given that E is the case, is p," be generalized to a theory of sentences of form "The probability of hypothesis H, given that the probability of E is q, is p"? Such a theory would seem to be necessary, e.g., if we are to cope adequately with the uncertainty attached to the background assumptions which always accompany a statistical analysis.

My suggestions for applied statistical analysis turn on the fact that while what is desired is the a posteriori probabilities of the various alternative hypotheses under consideration, computation of these by classical theory necessitates the corresponding a priori probability distribution, and in the more immediate future, at least, information about this will exist only as a subjective feel, differing from one person to the

next, about the credibilities of the various hypotheses.

3. Whenever possible, the basic statistical report should be in the form of a *confidence interval.* Briefly, a confidence interval is a subset of the alternative hypotheses computed from the experimental data in such a way that for a selected confidence level α, the probability that the true hypothesis is included in a set so obtained is α. Typically, an α-level confidence interval consists of those hypotheses under which the p value for the experimental outcome is larger than $1-\alpha$ (a feature of confidence intervals which is sometimes confused with their definition), in which case the confidence-interval report is similar to a simultaneous null-hypothesis significance test of each hypothesis in the total set of alternatives. Confidence intervals are the closest we can at present come to quantitative assessment of hypothesis-probabilities (see *technical note,* below), and are currently our most effective way to eliminate hypotheses from practical consideration—if we choose to act as though none of the hypotheses not included in a 95% confidence interval are correct, we stand only a 5% chance of error. (Note, moreover, that this probability of error pertains to the incorrect simultaneous "rejection" of a major part of the total set of alternative hypotheses, not just to the incorrect rejection of one as in the NHD method, and is a *total* likelihood of error, not just of Type I error.) The confidence interval is also a simple and effective way to convey that all-important statistical datum, the conditional probability (or probability density) function—i.e., the probability (probability density) of the observed outcome under each alternative hypothesis—since for a given kind of observed statistic and method of confidence-interval determination, there will be a fixed relation between the parameters of the confidence interval and those of the conditional probability (probability density) function, with the endpoints of the confidence interval typically marking the points at which the conditional probability (probability density) function sinks below a certain small value related to the parameter α. The confidence-interval report is not biased toward some favored hypothesis, as is the null-hypothesis significance test, but makes an impartial simultaneous evaluation of all the alternatives under consideration. Nor does the confidence interval involve an arbitrary decision as does the NHD test. Although one person may prefer to report, say, 95% confidence intervals while another favors 99% confidence intervals, there is no conflict here, for these are simply two ways to convey the same information. An experimental report can, with complete consistency and some benefit, simultaneously present several confidence intervals for the parameter being estimated. On the other hand, different choices of significance level in the NHD method is a clash of incompatible decisions, as attested by the fact that an NHD analysis which simultaneously presented two different significance levels would yield a logically inconsistent conclusion when the observed statistic has a value in the acceptance region of one significance level and in the rejection region of the other.

Technical note: One of the more important problems now confronting theoretical statistics is exploration and clarification of the relationships among inverse probabilities derived from confidence-interval theory, fiducial-probability theory (a special case of the former in which the estimator is a sufficient

statistic), and classical (i.e., Bayes') inverse-probability theory. While the interpretation of confidence intervals is tricky, it would be a mistake to conclude, as the cautionary remarks usually accompanying discussions of confidence intervals sometimes seem to imply, that the confidence-level α of a given confidence interval I should not really be construed as a probability that the true hypothesis, H, belongs to the set I. Nonetheless, if I is an α-level confidence interval, the probability that H belongs to I as computed by Bayes' theorem given an a priori probability distribution will, in general, *not* be equal to α, nor is the difference necessarily a small one—it is easy to construct examples where the a posteriori probability that H belongs to I is either 0 or 1. Obviously, when different techniques for computing the probability that H belongs to I yield such different answers, a reconciliation is demanded. In this instance, however, the apparent disagreement is largely if not entirely spurious, resulting from differences in the evidence relative to which the probability that H belongs to I is computed. And if this is, in fact, the correct explanation, then fiducial probability furnishes a partial solution to an outstanding difficulty in the Bayes' approach. A major weakness of the latter has always been the problem of what to assume for the a priori distribution when no pre-experimental information is available other than that supporting the background assumptions which delimit the set of hypotheses under consideration. The traditional assumption (made hesitantly by Bayes, less hesitantly by his successors) has been the "principle of insufficient reason," namely, that given no knowledge at all, all alternatives are equally likely. But not only is it difficult to give a convincing argument for this assumption, it does not even yield a unique a priori probability distribution over a continuum of alternative hypotheses, since there are many ways to express such a continuous set, and what is an equilikelihood a priori distribution under one of these does not necessarily transform into the same under another. Now, a fiducial probability distribution determined over a set of alternative hypotheses by an experimental observation is a measure of the likelihoods of these hypotheses relative to all the information contained in the experimental data, but based on no pre-experimental information beyond the background assumptions restricting the possibilities to this particular set of hypotheses. Therefore, it seems reasonable to postulate that the no-knowledge a priori distribution in classical inverse probability theory should be that distribution which, when experimental data capable of yielding a fiducial argument are now given, results in an a posteriori distribution identical with the corresponding fiducial distribution.

4. While a confidence-interval analysis treats all the alternative hypotheses with glacial impartiality, it nonetheless frequently occurs that our interest is focused on a certain selection from the set of possibilities. In such case, the statistical analysis should also report, when computable, the precise p value of the experimental outcome, or better, though less familiarly, the probability density at that outcome, under each of the major hypotheses; for these figures will permit an immediate judgement as to which of the hypotheses is most favored by the data. In fact, an even more interesting assessment of the postexperimental credibilities of the hypotheses is then possible through use of "likelihood ratios" if one is willing to put his pre-experimental feelings about their relative likelihoods into a quantitative estimate. For let $Pr(H,d)$, $Pr(d,H)$, and $Pr(H)$ be, respectively, the probability of a hypothesis H in light of the experimental data d (added to the information already available), the probability of data d under hypothesis H, and the pre-experimental (i.e., a priori) probability of H. Then for two alternative hypotheses H_0 and H_1, it follows by classical theory that

$$\frac{Pr(H_0, d)}{Pr(H_1, d)} = \frac{Pr(H_0)}{Pr(H_1)} \times \frac{Pr(d, H_0)}{Pr(d, H_1)} \quad [1][2]$$

[2] When the numbers of alternative hypotheses and possible experimental outcomes are transfinite, $Pr(d, H) = Pr(H, d) = Pr(H) = 0$ in most cases. If so, the probability ratios in Formula 1 are replaced with the corresponding probability-density ratios. It should be mentioned that this formula rather idealistically presupposes there to be no doubt about the correctness of the background statistical assumptions.

Therefore, if the experimental report includes the probability (or probability density) of the data under H_0 and H_1, respectively, and its reader can quantify his feelings about the relative pre-experimental merits of H_0 and H_1 (i.e., $Pr(H_0)/Pr(H_1)$), he can then determine the judgment he should make about the relative merits of H_0 and H_1 in light of these new data.

5. Finally, experimental journals should allow the researcher much more latitude in publishing his statistics in whichever form seems most insightful, especially those forms developed by the modern theory of estimates. In particular, the stranglehold that conventional null-hypothesis significance testing has clamped on publication standards must be broken. Currently justifiable inferential algorithm carries us only through computation of conditional probabilities; from there, it is for everyman's clinical judgment and methodological conscience to see him through to a final appraisal. Insistence that published data must have the biases of the NHD method built into the report, thus seducing the unwary reader into a perhaps highly inappropriate interpretation of the data, is a professional disservice of the first magnitude.

Summary

The traditional null-hypothesis significance-test method, more appropriately called "null-hypothesis decision [NHD] procedure," of statistical analysis is here vigorously excoriated for its inappropriateness as a method of *inference*. While a number of serious objections to the method are raised, its most basic error lies in mistaking the aim of a scientific investigation to be a *decision*, rather than a *cognitive* evaluation of propositions. It is further argued that the proper application of statistics to scientific inference is irrevocably committed to extensive consideration of inverse probabilities, and to further this end, certain suggestions are offered, both for the development of statistical theory and for more illuminating application of statistical analysis to empirical data.

REFERENCE

BRAITHWAITE, R. B. *Scientific explanation.* Cambridge, England: Cambridge Univer. Press, 1953.

(Received June 30, 1959)