In Praise of the Null Hypothesis Statistical Test

Richard L. Hagen
Florida State University

Jacob Cohen (1994) raised a number of questions about the logic and information value of the null hypothesis statistical test (NHST). Specifically, he suggested that: (a) The NHST does not tell us what we want to know; (b) the null hypothesis is always false; and (c) the NHST lacks logical integrity. It is the author's view that although there may be good reasons to give up the NHST, these particular points made by Cohen are not among those reasons. When addressing these points, the author also attempts to demonstrate the elegance and usefulness of the NHST.

In the past 30 years, a series of lively, imaginative articles have chided us for misinterpreting the null hypothesis statistical test (NHST; e.g., Brewer, 1985; Carver, 1978; Chronbach, 1975; Cohen, 1990, 1994; Meehl, 1967, 1978, 1990a, 1990b; Shaver, 1993; Snyder & Lawson, 1993; Thompson, 1993). In addition, these same articles have sometimes scolded us for trying to make statistical significance testing do what it was not meant to do, like appraising a grand theory on the basis of a few chi-squares (e.g., Meehl, 1990a).

Not infrequently, however, mixed in with these valuable admonitions, which we should certainly heed, have been criticisms of the usefulness and logical basis of the NHST. I believe these criticisms are not warranted. Of these articles, one by Cohen (1994) is particularly challenging and disturbing. Cohen presented some observations and "puzzles" that could lead a careless reader to the conclusion that statistical significance testing is worse than useless and should be abandoned. Although Cohen apparently did not come to this conclusion, some of his readers might.

An examination of Cohen's puzzles can lead us, I believe, to the opposite conclusion. It is within the context of three major points of disagreement with Cohen that I invite the reader to inspect these puzzles.

The NHST and the $P(H_0)$

Cohen (1994) stated that the NHST "does not tell us what we want to know," which is, "Given these data, what is the probability that H_0 is true?" (p. 997). He supported this point by providing a demonstration in which a Bayesian analysis seemingly led to a posterior probability of the null hypothesis quite different from the known true probability. His example suggests that after the data are in, the null hypothesis test can lead us further from the truth than we were before the experiment.

Apparently, this problem was no stranger to R. A. Fisher. Some of his correspondence suggests that he, like Cohen, also attempted to apply Bayesian theory to calculate the posterior probability of H_0, given the results of an experiment (Holschuh, 1980). The attempt did not work for Fisher any better than it did for Cohen.

But let's return to Cohen's example. In this example, the frequency of normal to schizophrenic individuals in some population is about 98 to 2. Accordingly, the probability of randomly drawing a normal individual is .98, and the probability of drawing a schizophrenic individual is .02. Because Cohen (1994) defined H_0 as "the case is normal" (p. 998), and H_1 as "the case is schizophrenic" (p. 998), it follows that the probability of H_0 is .98 and the probability of H_1 is .02.

Then, with a Bayesian analysis, Cohen established a posterior probability of .60, a number he called "the probability that the case is normal, given a positive test" (p. 999). The puzzle is that the .98 and the .60 both seem to refer to the probability that the null hypothesis is true. An even greater puzzle is that what appears to be the least accurate number is obtained after the researcher has attained significant data.

Is the example at fault? Or is the NHST flawed? I believe the problem lies with the example. Cohen tried to find a way to relate the probability of H_0 to "countable," "empirically based" relative frequencies. And his effort led him to define H_0 and H_1 in ways that the NHST does not allow.

Hayes (1963) reminded us that "A statistical hypothesis . . . is always a statement about the population, not about the sample" (p. 248). In Cohen's (1994) example, however, H_0 and H_1 are statements about the sample: "H_0 = the case is normal" and "H_1 = the case is schizophrenic" (p. 998).

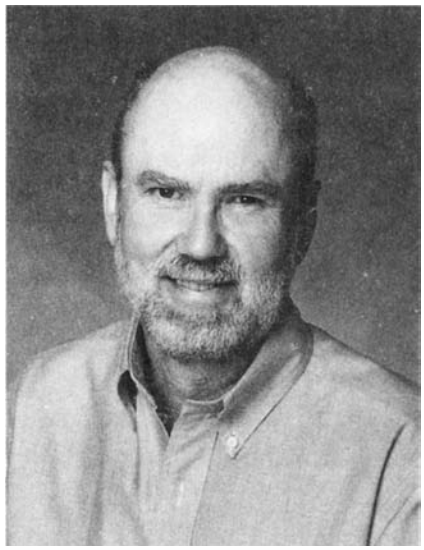
There are four issues raised by Cohen's choice to define H_0 and H_1 in this way:

Editor's note. J. Bruce Overmier served as action editor for this article.

Author's note. I am grateful for the constructive comments made by James K. Brewer, Judy Correa, Andrew Kaiser, Barbara Licht, Richard K. Wagner, and John Curtin. I am especially indebted to David Gustafson for his insightful contributions and to David Grunder for his tutoring on logical validity.

Correspondence concerning this article should be addressed to Richard L. Hagen, Department of Psychology, Florida State University, Tallahassee, FL 32306-1051. Electronic mail may be sent via Internet to hagen@psy.fsu.edu.

 0003-066X/97/$2.00
Vol. 52, No. 1, 15-24



Richard L. Hagen
Photo by Florida State University Photo Lab.

1. What is the population about which H_0 is supposed to make a statement?

2. Why cannot H_0 and H_1 in Cohen's example qualify as statements about such a population?

3. Why is it that H_0 and H_1 cannot be statements about a sample?

4. Why did Cohen break the rules?

First, the *population,* or sample space, is the set of possible outcomes for the experiment. In Cohen's example, the mixed pool of normal and schizophrenic individuals constituted that sample space. Therefore, H_0 must be a statement about this entire mixed group. So must H_1.

Second, the statement, "the case is normal" (H_0 in Cohen's example), is not a statement about a mixed population that contains both normals and schizophrenics. That statement can apply only to one segment of the population. A statement about the entire population—in this example, an acceptable null hypothesis—might be "the proportion of schizophrenics is less than 2%," or "exactly 2%," or something similar.

The reader might object: "But could we not decide to sample only from the group of normals or the group of schizophrenics? Maybe Cohen's H_0 can be saved." Yes, we could decide to sample from only one of these subgroups, but then the sample space, or population, becomes that subgroup, and it would not make sense for H_0 to say that the case is normal if we know we are sampling from a population of persons with schizophrenia.

We could, of course, draw samples from both of these populations. But the null still could not be about the sample. It might be a statement that embraces two parameters, one from each of the populations (e.g., the means of the two populations do not differ).

Perhaps there is a way to legitimize the null and the alternative hypotheses in Cohen's example, but I cannot find it.

The third question we might want to attend to is "Why must H_0 be a statement about a population?" Is that just a matter of convention? If it is, then Cohen's example is satisfactory. It is not just a matter of convention, and Cohen's example is very instructive in showing us why. There are two reasons that the null cannot be a statement about a sample:

1. If the null and the alternative hypotheses are statements about a sample, then the null hypothesis "becomes" true when, as in Cohen's example, we draw a case that is normal, and the alternative hypothesis "becomes" true when we draw a case that is schizophrenic. The sample, therefore, determines the status of the null hypothesis and, ultimately, the nature of our world. With the NHST, quite the opposite is true: It is the nature of our world that determines the characteristics of our sample.

This point can be brought home with greater clarity if one considers what the null hypothesis must be in a treatment outcome study. Suppose we are testing the effectiveness of a drug: H_0 = the drug does not work and H_1 = the drug does work. One of these hypotheses is a true statement about our world; the other is false. For the sake of this illustration, let us suppose that the drug really is effective (that is, the null is false).

We give a placebo to a control group and this effective drug to an experimental group. We then mix these participants into one group from which we select cases to test for prevalence of symptoms (like Cohen's mixture of normals and schizophrenics). Following the logic of Cohen's example, if we select by chance an experimental case, our selection means that the drug works, and if we select a control case, our selection means that the drug does not work. Surely, the world cannot be so capricious.

"Not so fast," the reader says. "Isn't the control sample, which did not receive the treatment, drawn from a world in which H_0 is true?" No, it is not. The untreated cases came from a world in which H_1 is true—*the treatment works in this world*—but it happens that the control sample was not subjected to the treatment.

As mentioned above, if the null and alternative hypotheses could be statements about samples, the particular sample we selected would determine the nature of our world. That argument alone should put to rest efforts to define H_0 and H_1 in terms of the sample. But consider what might happen if we drew two samples. We might end up with a world in which both the null and the alternative hypotheses are trying to be true at the same time. Such a world is not only capricious, it is impossible.

2. A statement about a sample cannot produce the required sampling distribution of the test statistic of interest.

I have criticized the way the example defines the null hypothesis. The burden is on me, therefore, to provide a better definition. What really is the null hypothesis?

At first glance, there appear to be many different null hypotheses: For example, the groups do not differ on a certain parameter, the correlation is zero, or the proportion of schizophrenics in such-and-such population is *P*. Is there a common attribute that ties together the many ways that the null hypothesis is stated?

At its core, the null hypothesis tells us something about a sampling distribution. A long form of the definition might be: The null states a condition, or set of conditions, in a possible world from which we can estimate the characteristics of a sampling distribution of a test statistic. It may be that those conditions involve taking a sample and using the characteristics of that sample plus a hypothesized parameter to estimate the sampling distribution. It may be that the conditions involve only a stated proportion and a formula, as in the case of the binomial distribution. But the core is always the same: The null provides a "theoretically knowable" sampling distribution against which we can compare our obtained test statistic to see how unusual our statistic would be if it were produced under the null. The statistical tables in the back of statistics books provide us with a few specks of that sampling distribution. Those few specks are our critical values.

By contrast, the alternative, or experimental hypothesis, states a condition in a possible world that does not tell us the characteristics of a sampling distribution (and accordingly, there are no tables in the back of statistics books against which we can test the alternative hypothesis).[1] The nature of the sampling distribution that would be produced under the alternative is unknowable because we can never know the true population parameter when H_1 is true (e.g., the mean of the treated group, the size of the true correlation, the true relative frequency, etc.).

Students sometimes wonder why it is that we test the null when it is the experimental hypothesis we really are interested in. The above definitions of the null and alternative hypotheses make the answer clear: We only test the null because only under the null can we know what the sampling distribution of the test statistic is supposed to look like.

Now back to Cohen's (1994) example. The null hypothesis stated that the case is normal. This statement about a sample cannot tell us anything about a sampling distribution of a test statistic. Accordingly, it is not an acceptable statement of a null hypothesis.

The fourth question raised above is, "Why did Cohen devise an example that breaks the rules?" For his example to work, he had to have countable, relative frequencies of H_0 and H_1. And for these frequencies to be relative, H_0 and H_1 had to be related to events that exist in the same sample space, or population. Because neither H_0 nor H_1 could make a statement about a population that included both H_0 and H_1 events, the only alternative was to have H_0 and H_1 make statements about the sample.

I suspect that we cannot construct an example that meets the requirements of the NHST and also provides countable frequencies. If this is true, we will never be able to use a counting method of probability in a Bayesian analysis to establish the posterior probabilities of H_0 and H_1. From what Fisher wrote, it appears that this is the reason he gave up the attempt some 60 years ago.

What good, then, is the NHST? And how can it give us any meaningful information about the probability of H_0? Should we abandon the model? Fisher did not. Instead, he suggested that we reject the use of Bayes' theorem in trying to make inferences about the posterior probability of H_0 (Holschuh, 1980; Lane, 1980).

Yet, Cohen (1994) rightly argued that the posterior probability is available only through Bayes' theorem, and indeed, it is that posterior probability that "we so much want to know" (p. 997). Can we resolve this dilemma?

The $P(H_0)$ as a Subjective Level of Confidence

It is in Fisher's (1937) comments on why he rejected a Bayesian analysis that we see what may be a solution. Fisher's major complaint was that Bayesian reasoning requires one to

> regard mathematical probability not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes. (p. 6)

Fisher was saying that frequency-derived (objective) probabilities will not work in a Bayesian computation of the probability of the null and that the only approach that will work—treating probabilities as equivalent to subjective levels of confidence—is unscientific. Fisher did not embrace this "unscientific" solution for many years; instead, he continued for most of his life to insist that probability statements are statements about frequencies of attributes in populations, never about levels of beliefs (Lane, 1980).

His position did not go unopposed. A major antagonist over the nature of probabilities was Sir Harold Jeffreys who, through a series of sometimes heated exchanges with Fisher between 1932 and 1934, argued for a subjective and psychological definition of probability. (See the *Proceedings of the Royal Society of London,* Vol. 138–146). In regard specifically to the prior probabilities one uses in a Bayesian analysis, Jeffreys insisted that "A prior probability is not a statement about frequency of occurrence in the world or in any portion of it" (Jeffreys, 1934, p. 9). Accordingly, a Bayesian posterior probability, which is a function of Bayesian priors, also could not be a statement about frequency.

Now, let us return to our question: Does the NHST tell us what we want to know about the $P(H_0)$? The critical issue in answering this question is how we define

[1] With a power analysis, what is called the alternative hypothesis does provide a sampling distribution to which the obtained statistic can be compared. As pointed out later in this article, a power analysis essentially converts the alternative hypothesis to a null hypothesis, which can then be rejected the same way any null hypothesis can be rejected.

probability. Recognizing that there are a number of different ways that probability may be defined (cf. Carnap, 1945), we need only to focus on two of these to demonstrate what the NHST does and does not tell us.

As mentioned above, throughout most of his life, Fisher preferred to define probability in terms of observable frequencies. By contrast, Jeffreys, following the lead of the Reverend Thomas Bayes some two centuries earlier, defined probability as a degree of belief (Earman, 1992). This same concept, essentially unchanged, is represented by a number of modern writers who claim that a Bayesian posterior probability can only be construed as a subjective "level of confidence" or "estimate of likelihood" (e.g., Darnell & Evans, 1990; Klayman & Ha, 1987; Savage, 1954, 1962). This construal of probability is familiar to all of us in the form of a statement of *odds,* a concept which Bayes himself used when he attempted to explain what he meant by degree of belief.

Both the "relative-frequency" approach to probability (Fisher's [1937] preference) and the "degree-of-belief" approach (Jeffreys' [1934] preference) can be said to have logical validity (Carnap, 1945). Therefore, the issue in applying one or the other to statistical significance testing is not which is correct but which seems to work.

I invite those who have some comfort with Jeffreys' approach to now return with me to the questions Cohen (1994) raised about the NHST. Does the NHST assist us in making decisions about the nature of the world? Does the NHST increase or decrease our subjective levels of confidence about the existence of phenomena we are attempting to study? Does the NHST tell us anything at all about what we want to know?

Let us use Cohen's example of psychiatric diagnosis to demonstrate what the NHST can tell us. In this example, 98% of a population are normals and 2% are schizophrenics. A test for schizophrenia identifies 97% of the normals and 95% of the schizophrenics. In this example, therefore, the following probabilities are obtained:

Probabilities based on relative frequencies	**In Cohen's formula**
P(drawing a normal individual) = .98	$P(H_0$ is true)
P(drawing a schizophrenic individual) = .02	$P(H_1$ is true)
P(positive test | case is normal) = .03	Alpha
P(negative test | case is schizophrenic) = .05	Beta

The formula, then, for the Bayesian posterior probability that the case is normal, given a positive test, is calculated in the following way (see Appendix for an explanation of how the formula was derived):

$$P(H_0 \mid \text{positive test}) = \frac{P(H_0)(\text{alpha})}{P(H_0)(\text{alpha}) + P(H_1)(1 - \text{beta})}$$

$$= \frac{(.98)(.03)}{(.98)(.03) + (.02)(.95)} = .607.$$

According to Cohen (1994), this number, .60, "demonstrates how wrong one can be by considering the p value from a typical significance test as bearing on the truth of the null hypothesis for a set of data" (p. 999). He is correct if we try to conceptualize the $P(H_0)$ in terms of a countable, empirically based relative frequency. If, on the other hand, we think of the $P(H_0)$ as a degree of belief, or level of confidence, his example demonstrates how much the Bayesian posterior probability can tell us about the $P(H_0)$.[2]

Instead of having the .98 represent the relative frequency of normals, let us think of it as representing the researcher's very high level of confidence that a treatment will not work (or that a condition does not exist). Conversely, the researcher's hunch is that there are only about 2 chances in 100 that the treatment will work.

According to Cohen's arithmetic, with which I concur, after obtaining significant results, the researcher's level of confidence that there is no treatment effect has decreased from .98 to about .60, and level of confidence that the treatment did have an effect has increased from .02 to about .40.

In this particular example, Cohen favored us with power = .95, which makes a replication highly likely if, indeed, H_1 is true. And given the dismal odds the researcher originally allowed, an attempt to replicate would be in order.

With one replication, using the level of confidence derived from Experiment 1 (and the same alpha and beta that Cohen used), a Bayesian analysis shows that the prudent researcher should reject H_0.

$$\text{Subjective } P(H_0 \mid \text{positive test}) = \frac{(.03)(.60)}{(.03)(.60) + (.95)(.40)} = .045.$$

The researcher's level of confidence that there is no treatment effect is now down to .045, and the researcher can

[2] The example is instructive about the usefulness of a conditional probability apart from any implications having to do with the probability of a null hypothesis. As Cohen (1994) pointed out, the Bayesian posterior probability of .60 is the expected relative frequency of false positives among all cases that have been diagnosed as schizophrenic from the test results. Accordingly, .40 is the expected relative frequency of true positives among those who score positive. The .40 represents a dramatic increase over base rate in the probability of correctly identifying schizophrenics, who, in the population at large, compose only 2% of the cases. A similar analysis shows that among those who score negative, the probability of misdiagnosis is less than .001. If the figures in this example were applied to a screening test for a serious illness, the further diagnosis of which involved intrusive, expensive tests, the screening test would be of considerable value.

say, "Based on my original estimate of the likelihood that the treatment would work (which was very low) and twice obtaining significant data, I now estimate that the odds that this treatment does work are about 21 to 1."

The example Cohen chose has some characteristics that may not be typical of the average research problem. First, the researcher's initial level of confidence in H_0 was unusually high (.98); second, power was an almost unheard of .95. Will a Bayesian calculation of the inverse subjective probability of H_0 work with less extreme numbers? Yes, it will.

Consider the following example. An experimenter believes that the odds are 50–50 that a treatment will have an effect. Let us assume $\alpha = .05$ and power = .40 (about what Cohen [1962] found in his survey of articles in the *Journal of Abnormal and Social Psychology*). The Bayesian formula for calculating the subjective probability of H_0, given data that fall into the alpha region (symbolized by D*), is the following:

$$\text{Subjective } P(H_0|D^*) = \frac{(.05)(.5)}{(.05)(.5) + (.4)(.5)} = .11.$$

From these figures, we can see that after obtaining a significant effect, the researcher's initial degree of confidence that the treatment would not work is decreased from .50 to .11. Conversely, the degree of confidence that the treatment did work is increased from .50 to .89.

If one uses an alpha level of .01 for this same problem, degree of confidence in the treatment increases from .50 to .98 after a significant result is obtained.

The above examples show, and others would continue to show, that when D* is obtained, the researcher's level of confidence that the treatment was effective is always increased. What happens to level of confidence when the data obtained do not fall in the range of scores specified by alpha (when D, not D*, is obtained)?

A Bayesian analysis using the same formula as above but calculating the probability of H_0 given D, instead of D*, shows that a failure to obtain significant results increases one's level of confidence in H_0. Why then, do we not accept H_0? Why do most of our textbooks insist that H_0 can only be rejected and cannot be accepted? Cohen (1990) framed this puzzle in the following way:

> Another problem that bothered me was the asymmetry of the Fisherian scheme: If your test exceeded a critical value, you could conclude, subject to the alpha risk, that your null was false, but if you fell short of that critical value, you couldn't conclude that the null was true. In fact, all you could conclude is that you couldn't conclude the null was false. In other words, you could hardly conclude anything. (p. 1308)

The scheme is asymmetrical because the increases in levels of confidence in H_0 and H_1, given D and D* respectively, are almost always asymmetrical. They would only be symmetrical when the subjective priors of H_0 and H_1 are equal and when $\alpha = \beta$, a condition that may occur in the real world but one that is usually too expensive for the researcher to set up for an a priori power analysis.

Figure 1 shows the asymmetry when the subjective prior probability of H_0 = .50 and power = .40. As seen in these functions, level of confidence in H_0 is increased by a failure to obtain D*, but that increase in confidence in H_0 is small compared with the increase in confidence in H_1 when D* is obtained.

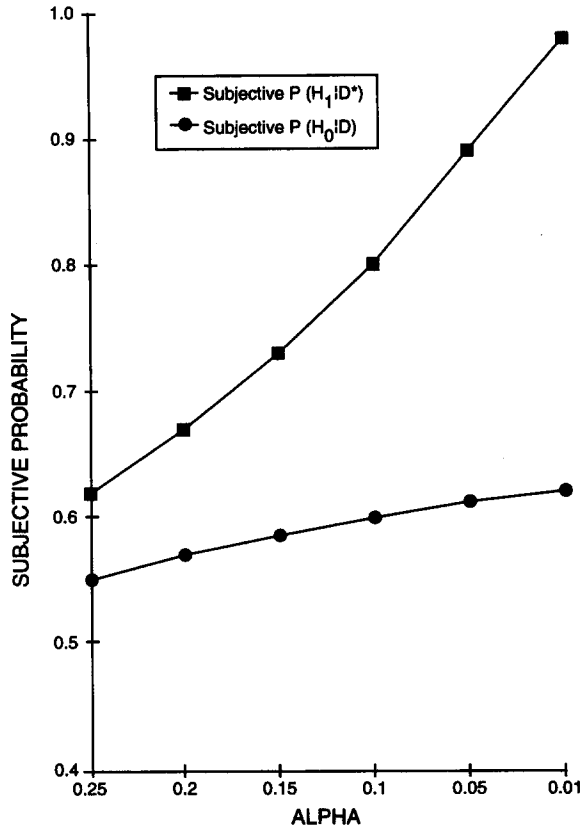
Cohen (1994) stated that the NHST "does not tell us what we want to know," which is "Given these data, what is the probability that H_0 is true?" (p. 997). If one is seeking a frequency-based probability, he is correct. But if we are content to equate the $P(H_0)$ with a subjective degree of belief, or level of confidence, then the NHST does, indeed, tell us what we want to know.

The NHST and the Status of the Null

> So if the null hypothesis is always false, what's the big deal about rejecting it? (Cohen, 1994, p. 1000)

Cohen's comment is reminiscent of Meehl's (1978) claim that "the null hypothesis, taken literally, is always false," that this fact is "generally recognized by statisticians

Figure 1
Subjective Probabilities of $H_0|D$ and $H_1|D^*$ as a Function of Alpha When the Prior Subjective Probabilities of H_0 and H_1 = .5



today and by thoughtful social scientists," and "that among sophisticated persons, it is taken for granted" (p. 822).

The contexts of these quotations from Cohen and Meehl suggest that they were talking only about "soft psychology," which apparently refers to (a) the study of variables, each set of which comes from the same individual (or same entity), or (b) the study of differences among intact groups. Under either of these two conditions, the null hypothesis may always be false. When scores are pulled from the same pocket, they probably are related to each other for the very reason that they fell into that pocket to begin with. In addition, it is unlikely that any two intact groups, particularly if those groups have quite different histories, are exactly the same on any variable one might measure.

Meehl (1990b) made clear in a later article that he did not mean that the null hypothesis is always false in "purely experimental studies" (p. 204). In addition, in that same article, even when speaking of soft psychology, Meehl was careful to state that the null hypothesis is "almost" (p. 124) always false. I believe it is safe to assume that Cohen would agree with Meehl and that when Cohen (1994) said, "the null is always false" (p. 1000), he only wanted to rattle us into more careful thinking.

So then, what is the problem? The problem is that an uncritical reader may assume what these writers (Cohen, 1990, 1994; Meehl, 1990b) did not intend. During the past few years, several colleagues have referred me to Cohen's 1990 article as they have tried to convince me that the null is always false under all circumstances. Cohen may not believe it, but apparently some of his readers do.

On what basis would one come to the conclusion that the null hypothesis is always false? The argument that appears to be most compelling is this: If the measure is fine enough, any observed samples will differ on whatever variable we might choose to measure, and, therefore, the null hypothesis, taken literally, is always false. This appears to be the argument Cohen (1990) used in the following statement:

> A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally . . . is *always* [Cohen's italics] false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). (p. 1308)

In this statement, Cohen apparently was saying that the null is false when *samples* are unequal on some variable. His comment suggests that he believes that even if a computer were programmed to make the null true by generating samples that were exactly equal, a stray electron might make the null false by producing samples that differed is some slight way.

But the null hypothesis says nothing about samples being equal, nor does the alternative hypothesis say that they are different. Rather, when addressing group differences, the null hypothesis says that the observed samples, given their differences, were drawn from the same population, and the alternative hypothesis says that they were drawn from different populations.

Samples drawn from the sample population will always differ. No matter how "randomly" participants are assigned, the groups will differ in an absolute sense on any measurable variable if the measure is fine enough. They will differ both before and after the experimental procedures are carried out. Groups drawn from the same population would be equal on a variable only under sampling with replacement and only if the sample size was equal to the population size. Fisher was surely aware of this. The system he and his followers devised not only anticipates the presence of such differences but it also accommodates them in the span of "$1 - \alpha$" in the sampling distribution of the statistic used to test the null.

A second argument that has been put forth to support the idea that the null is always false is this: Even when samples are drawn from the same population, the absolute differences between samples will always reach significance given a large enough N. We have been taught that a sufficiently large N will detect differences no matter how tiny they may be. But what we may forget is that small differences will always be detected by a large N only under the alternative hypothesis, not under the null.

When samples are drawn from the same population, the variance of absolute differences between or among such samples will become smaller as N becomes larger. This diminishing variance is reflected in a decrease in the variance of the particular test statistic from which we draw our sample statistic. Accordingly, Type I error remains roughly constant no matter how large N becomes. Thus, although it may appear that larger and larger Ns are chasing smaller and smaller differences, when the null is true, the variance of the test statistic, which is doing the chasing, is a function of the variance of the differences it is chasing. Thus, the "chaser" never gets any closer to the "chasee."

A third argument is based on the assumption that whenever groups are treated differently in any way, those different treatments will always have some differential effect on the groups. By quoting Tukey (1991), Cohen (1994) suggested that this argument relates to the possibility that the null is always false: "It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place" (Tukey, 1991, p. 100).

"A and B" can refer not only to the experimental manipulation but also to the countless other ways—often unintended and unnoticed—in which the groups will be treated differently. It may seem reasonable, therefore, to assume that such differences, which are more than just sampling error, would always lead to a rejection of the null hypothesis if N is sufficiently large.

I agree that A and B will always produce differential effects on some variable or variables that theoretically could be measured. But I do not agree that A and B will always produce an effect on the dependent variable (DV), and it is, after all, only a difference on the DV that will

lead to a rejection of the null hypothesis. Consider the theoretical leaps we would have to make if we were to accept the idea that all differences in treatment would produce changes that would ultimately show up as differences on the DV. A few years ago, visual imagery therapists were treating AIDS patients by asking the patients to imagine little AIDS viruses in their bodies being eaten by monsters. Under such a treatment, both psychological and physiological changes would take place ("No twisted thought without a twisted neuron," Karl Lashley supposedly said). Thus, Tukey (1991) was correct: The effects of A and B are different. But many would question whether or not such changes would be reflected in the participant's T-cell count. And if that is the DV, it is only that difference that would lead to a rejection of the null hypothesis.

Or consider an experiment in which there is a large number of unintended differences in treatment. Suppose we attempt a sloppy replication of a study Johnson and Barber (1978) conducted years ago on the effects of hypnosis on warts. Let's suppose furthermore that the treatments are carried out in different rooms, with different kinds of chairs, different lighting, perhaps even different wallpaper. For these uncontrolled and unplanned differences to be reflected on the dependent measure, they would have to be related to differential disappearance of warts. And in that case, forget the hypnosis, we might make a fortune selling wallpaper to witches.

The point is that groups can be treated differentially in many ways, but those differences may not relate conceptually or empirically to the DV. Tukey's (1991) comment that the effects of A and B are always different can stand. But it does not necessarily follow that the null hypothesis will always be vulnerable to those effects.

The final argument I wish to address makes the following claim: If an experiment is repeated enough times, eventually H_0 will be rejected. Therefore, H_0 can always be shown to be false.

I agree wholeheartedly with the first part of this assumption. If an experiment is repeated enough times, eventually H_0 will be rejected for a given experiment, even if H_0 is true. But it does not follow that H_0 can therefore always be shown to be false. If we flipped a coin seven times, we would not expect to obtain seven heads on those first seven flips. However, if we flipped the coin seven more times, then seven more times, then seven more times, eventually, on one of these sets of seven flips, we would obtain seven heads in a row.

If those seven heads in a row came after 100 or so sets of flips, we would not conclude that the coin was biased (that is, we would not reject the null). Instead, we would consider all the data and would conclude that the one run of seven heads was to be expected occasionally under the null. This approach is commonly used in the interpretation of a large number of correlations, of which a few are significant. The NHST is not embarrassed by demonstrations that Type I errors can be produced given a large number of replications.

If, as some have claimed, the null hypothesis is always false, we would be foolish, indeed, to spend time conducting statistical tests that can only tell us what we already know. But we need not feel foolish. As far as I can tell, the claim has never been sustained by either statistical or logical arguments.

The NHST and Formal Logic

The logic on which the NHST rests has come under regular fire during the past 35 years (e.g., Carver, 1978, 1993; Rozeboom, 1960; Shaver, 1993; Thompson, 1993). Criticisms have ranged from calling the logic tautological to challenging the backwardness of having to assume something is true to demonstrate that it is false. A full consideration of the logical basis of the NHST is far beyond the scope of this article. I will focus on Cohen's (1994) critique of the NHST logic, a critique that, to my knowledge, has not been raised before.

Cohen (1994) cautioned us not to be seduced into attributing logical validity to the NHST. The steps in his argument are the following:

1. The NHST syllogism is similar in form to modus tollens.

2. The modus tollens form has logical validity; therefore, people might be tempted to think that the reasoning of the NHST also has logical validity.

3. But when the modus tollens form is stated probabilistically, the form is no longer logically valid.

4. Therefore, in spite of what might appear to be the case, the NHST syllogism does not have formal logical validity.

Cohen stopped there, and the reader must draw his or her own conclusion as to whether or not this means that the NHST form is illogical.

My lack of sophistication in formal logic led me to the conclusion that Cohen was, indeed, implying that the NHST form is illogical, but a colleague, a professor of philosophy, suggested that Cohen was probably only pointing out that people are misled into thinking that the NHST has formal validity because it is so similar to modus tollens. This colleague also alerted me to several other characteristics of formal logic, characteristics that I use in the following reply to Cohen's criticism of the logic of the NHST:

1. Only certain forms of reasoning (e.g., modus ponens, modus tollens) are accorded formal logical validity.

2. If an argument is presented in one of these forms, it is always valid (has formal logical validity), but it is not always sound.

Example:

If you contract AIDS, you will be healthy and happy.
You did contract AIDS.
You are healthy and happy.

This argument is logically valid. We might wonder, therefore, how much importance we should attach to logical validity as a criterion for scientific argument.

3. On the other hand, arguments can be reasonable and defensible even when they are not logically valid in a formal sense.

Example:

If you contract AIDS, you will probably die of some opportunistic infection within 10 years.
You did contract AIDS.
You will probably die of some opportunistic infection within 10 years.

This probabilistic argument is not formally logical because one could accept the premises but still reject the conclusion. The argument is, however, quite reasonable and defensible based on data.

Most of the decisions we make throughout our lives are based on probabilistic premises, not on logic that is valid in a formal sense. The conclusions I draw in this article concerning the usefulness of the NHST are not logically valid in any formal sense. Nor are our suspicions concerning the faithfulness of our spouses or the competence of our auto mechanics. Our court system would have to close down if only formal logic were allowed. No conclusions derived from an argument stating, "The evidence suggests that . . . ," are ever logically valid.

Science has done well using arguments that are not logically valid. In the absence of an alternative, it will have to continue to do so.

Not only will scientists have to continue using arguments that lack formal logical validity, but they may have a difficult time getting rid of the seemingly backward logic that is peculiar to methods of statistical significance testing. For example, point estimates and confidence intervals have much to offer (see Cohen, 1994; Schmidt, 1996), and at first glance, it appears that with a confidence interval we might be able to test a hypothesis concerning a parameter in a straightforward manner. After all, doesn't a confidence interval tell us the range in which we *should* find the parameter rather than the range in which we *should not* find it? Is that not a way to avoid the logic imposed on us by the NHST?

Sorry, but it is not. The same logical form is there. But instead of starting with a hypothesized parameter and from that parameter developing a sampling distribution against which our sample statistic would be compared, we begin with the statistic, establish a confidence interval, and against that confidence interval we test an infinite number of parameters. The upper and lower limits of the confidence interval are defined by hypothesized values of the parameter in question that would be rejected right at the .05 level by the data, and, accordingly, all hypothesized values of the parameter below the lower limit or above the upper limit would be rejected. Hypothesized values of the parameter that fall within the confidence interval would not be rejected.

For a parameter outside of the confidence interval to be rejected, one must invoke the following logic: The probability that a population with "this" parameter produced "this" datum from which "this" confidence interval was constructed is very low. Therefore, we reject the idea that the datum came from such a population. We cannot escape the logic of the NHST by turning to point estimates and confidence intervals.

Similarly, we cannot avoid this logic by doing a power analysis. When we specify power, we state what the probability should have been of detecting at least a certain effect size. In a test for a mean difference, for example, we have, in effect, created a "new" null hypothesis which posits a new null distribution, the stated μ of which is the effect size away from the μ of the old null distribution. This new null hypothesis meets all the requirements of a null mentioned above because with it, we can specify a theoretical sampling distribution of means against which we can compare our obtained mean. Similarly, the old null hypothesis now loses its identity with a stated μ: It can no longer tell us about a theoretical sampling distribution of means, and, therefore, it no longer meets the requirements of a null hypothesis. What has happened is that the original null and alternative hypotheses have switched places, and what formerly was beta is now the new alpha. As long as we do power analyses, we are stuck with the logic of the NHST.

At its simplest level, the NHST logic is used to evaluate the significance of a two-variable correlation or a difference between two groups. With more complex inferential methods, the same logic is used to evaluate an F change in multiple regression, a departure from 1.00 in proportion of variance among effect sizes (Feingold, 1995), or the fit of a covariance structure model to the obtained data.

The logic of the NHST is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want to.

As mentioned at the outset of this article, the NHST has been misinterpreted and misused for decades. This is our fault, not the fault of the NHST. I have tried to point out that the NHST has been unfairly maligned; that it does, indeed, give us useful information; and that the logic underlying statistical significance testing has not yet been successfully challenged. Given the controversy that continues to reign over the NHST, what shall we do? Cohen (1994) asked this question in the article to which my reply has been directed. His answer was, "Don't look for a magic alternative to the NHST . . . It doesn't exist" (p. 1001). I suspect he was right. I would add that, when we use the NHST, as I suspect most of us will continue to do, let's not forget to celebrate its brilliance once in a while.

REFERENCES

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10,* 252–268.

Carnap, R. (1945). The two concepts of probability. *Philosophy and Phenomenological Research, 5,* 513–532.
Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.
Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287–292.
Chronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.
Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 69,* 145–153.
Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.
Darnell, A. C., & Evans, L. (1990). *The limits of econometrics.* London: Aldershot.
Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory.* Cambridge, MA: MIT Press.
Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist, 50,* 5–13.
Fisher, R. A. (1937). The design of experiments. London: Oliver & Boyd.
Hayes, W. L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart & Winston.
Holschuh, N. (1980). Randomization and design: I. In S. Fienberg, J. Gani, J. Kiefer, & K. Krickeberg (Eds.), *R. A. Fisher: An appreciation* (pp. 35–84). New York: Springer-Verlag.
Jeffreys, H. (1934). Probability and scientific method. *Proceedings of the Royal Society of London, Series A, 146,* 9–15.
Johnson, R. F., & Barber, T. X. (1978). Hypnosis, suggestion, and wants: An experimental investigation implicating the importance of "believed in" efficacy. *American Journal of Clinical Hypnosis, 20*(3), 165–174.
Klayman, J., & Ha, Y. M. (1987). Confirmation, disconformation, and information in hypothesis testing. *Psychological Bulletin, 94,* 211–228.
Lane, D. A. (1980). Fisher, Jeffreys, and the nature of probability. In S. Fienberg, J. Gani, J. Kiefer, & K. Krickeberg (Eds.), *R. A. Fisher: An appreciation* (pp. 148–160). New York: Springer-Verlag.
Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.
Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.
Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.
Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.
Savage, L. J. (1962). Subjective probability and statistical practice. In G. A. Barnard & D. R. Cox (Eds.), *The foundations of statistical inference: A discussion* (pp. 9–35). New York: Wiley.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.
Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61,* 293–316.
Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61,* 334–349.
Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Research, 61,* 361–377.
Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

(Appendix follows on next page)

Appendix

Explanation of Formula for Bayesian Posterior Probability of H_0

The density functions below represent sampling distributions from two possible worlds. In one world, H_0 is true; in the other, H_1 is true. Although the two sampling distributions cannot exist in the same world at the same time, when represented as "possible," they can be placed above a common scale of measurement.

A critical value divides the H_0 distribution into two parts: α and $1 - \alpha$ (determined by the experimenter). That same critical value divides the H_1 distribution into β and $1 - \beta$ (determined by N, the actual effect size, and the reliability of the measures). To the right of the critical value, lie significant values of the test statistic (D*); to the left lie nonsignificant values (D).

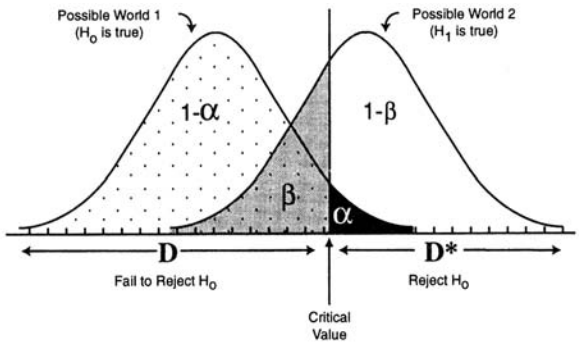
When the probabilities of H_0 and H_1 are equal, the probability that our data (the sample statistic) came from a particular distribution is:

$$\frac{\text{the sample space of that distribution}}{\text{the total sample space represented by both distributions}}.$$

For example, when significant data (D*) are obtained, D* can have come only from the α portion of the H_0 distribution or from the $1 - \beta$ portion of the H_1 distribution. Accordingly, the

$$P(H_0|D^*) = \frac{\alpha}{\alpha + (1 - \beta)}.$$

Similarly, nonsignificant data can only be obtained from either the $1 - \alpha$ portion of the H_0 distribution or the beta portion of the H_1 distribution. Accordingly, the



$$P(H_0|D) = \frac{(1 - \alpha)}{(1 - \alpha) + \beta}.$$

Using the same approach, one can calculate the probability of H_1 given D* or D.

When the prior probabilities of H_0 and H_1 are not equal—that is, when the subjective judgment of the experimenter places a higher probability on one possible world than on the other—then the prior probability attached to each possible world must enter into the equation for the computation of a Bayesian posterior probability. For example,

$$\text{Posterior } P(H_0|D^*) = \frac{[\text{Prior } P(H_0)][\alpha]}{[\text{Prior } P(H_0)][\alpha] + [\text{Prior } P(H_1)][1 - \beta]}.$$