American Journal of
EPIDEMIOLOGY

Volume 137
Number 5
March 1, 1993

Sponsored by the Society for Epidemiologic Research

REVIEWS AND COMMENTARY

p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate

Steven N. Goodman

It is not generally appreciated that the *p* value, as conceived by R. A. Fisher, is not compatible with the Neyman-Pearson hypothesis test in which it has become embedded. The *p* value was meant to be a flexible inferential measure, whereas the hypothesis test was a rule for behavior, not inference. The combination of the two methods has led to a reinterpretation of the *p* value simultaneously as an "observed error rate" and as a measure of evidence. Both of these interpretations are problematic, and their combination has obscured the important differences between Neyman and Fisher on the nature of the scientific method and inhibited our understanding of the philosophic implications of the basic methods in use today. An analysis using another method promoted by Fisher, mathematical likelihood, shows that the *p* value substantially overstates the evidence against the null hypothesis. Likelihood makes clearer the distinction between error rates and inferential evidence and is a quantitative tool for expressing evidential strength that is more appropriate for the purposes of epidemiology than the *p* value. *Am J Epidemiol* 1993;137:485–96.

hypothesis tests; inference; likelihood; *p* values; significance tests

Editor's note: *For a discussion of this paper and for the author's response, see pages 497 and 500, respectively.*

Much has been written about the proper roles for inductive and deductive reasoning in epidemiology (1–3), but it is unclear to what degree such discussions affect epidemiologic practice. In this essay, we will review a historical debate about induction and deduction that has direct relevance to the way epidemiologists express quantitative uncertainty and to the methods they use daily. It is a debate that continues at a lower ebb than in the past because it has suffered a curious fate: The approaches of each side have been improperly combined, creating a new procedure with such a strong illusion of coherence that even when it produces problems, the combination is not recognized as their source.

This "new procedure" is the one currently

Received for publication November 4, 1991, and in final form August 5, 1992.

From the Division of Biostatistics, Oncology Center, The Johns Hopkins University School of Medicine, Baltimore, MD.

Reprint requests to Dr. Steven N. Goodman, 550 N. Broadway, Suite 1103, Baltimore, MD 21205.

The author thanks Dr. Sander Greenland, Dr. Steven Lane, and Dr. Teddy Seidenfeld for helpful suggestions on this paper and Dr. Richard Royall for introducing him to the subject.

used in most epidemiologic and medical research. An experiment is designed to control the probabilities of two types of "error," designated type I (α, usually equal to 0.05) and type II (β, usually less than 0.20). When the data are in, a p value is used as a quantitative measure of evidence against the null hypothesis. If the p value is less than α, the result is declared "significant," and the null hypothesis is regarded as unlikely to be true.

Numerous writers have pointed out the dangers of drawing automatic conclusions based on this procedure (4–8), the importance of using biologic judgment to interpret such results, and its problems from a Bayesian or likelihood perspective (9–17). In many institutions, those caveats are part of standard epidemiologic teaching. What we will examine here is a conflict within the method itself: the incompatibility of the p value with the hypothesis test in which it is today imbedded. We will then show how likelihood methods can both illuminate and resolve the conflict. Though this situation has been pointed out by a number of statisticians and philosophers (18–28), most researchers are unaware of its existence, no less its implications. It has become so blurred over the decades that for clarity we need to return to the time and writings of the scientists who developed the original ideas: R. A. Fisher for the p value and Jerzy Neyman for the hypothesis test.

R. A. FISHER AND THE p VALUE

Some features of R. A. Fisher's life provide insight into his scientific philosophy, which has a direct bearing on the interpretation of the p value. R. A. Fisher, who has been called the "father of modern statistics," was interested in the fields of biometry and genetics, and he played a major role in bringing both fields out of their infancies. Though he was an abstract thinker of the highest order, Fisher regarded with scorn solutions to biologic problems derived without a full understanding of the reasoning used by the experimenters. He stated that the teaching of statistics should be entrusted "only to such mathematicians as have had sufficiently prolonged experience of practical research, and of responsibility for drawing conclusions from actual data, upon which practical action is to be taken" (29, p. 435). He ultimately held professorships in eugenics and genetics, but never had a permanent academic appointment in statistics.

As a practicing scientist, Fisher had an abiding interest in creating an objective, quantitative method to aid the process of inductive inference—drawing conclusions from observations. He did not believe that the use of the Bayes formula to convert prior probabilities of hypotheses (before the data) to posterior probabilities (after the data) was justified in scientific research, where prior probabilities are usually uncertain. He ultimately proposed three inferential methods that did not require prior probabilities of hypotheses. The first two were relatively informal, one based on the p value and the other on mathematical likelihood. A third, formal method was called "fiducial inference." It was generally regarded as unsuccessful and will not be discussed here.

Fisher was not the first to use the p value, but he was the first to outline formally the logic behind its use, as well as the means to calculate it in a wide variety of situations. Fisher's definition for the p value, or "significance probability," was essentially the same used today: it equaled the probability of a given experimental observation, plus more extreme ones, under a null hypothesis. If this number were small, one could "reject" the null hypothesis as unlikely to be true. The use of a threshold p value as a basis for rejection was called a "significance test." This is important to distinguish from the "hypothesis test," which will be discussed shortly.

Fisher's formal definition of the p value appears similar to the one used today, but his notions about how it was to be used and interpreted were somewhat different. First, the p value was not to be interpreted as a hypothetical frequency of "error" if the experiment were repeated. It was a measure of evidence in a single experiment, to be used to reflect on the credibility of the null hypothesis, in light of the data. Second, as a

measure of evidence, the p value was meant to be combined with other sources of information about the phenomenon under study. If a threshold for "significance" was used, it was to be flexible and to depend on background knowledge about the phenomenon being studied (30).

The idea that a p value should be used flexibly as a measure of evidence within a complex descriptive and inferential process is shared by most epidemiologists today. But Fisher's rejection of its frequency interpretation is contrary to how most modern researchers conceive of it. If the p value does not reflect the frequency of hypothetical results upon repetition of the experiment, how can its numerical value be interpreted? This was not the only unanswered question implicit in Fisher's significance probability. Which outcomes were "more extreme," and how were they relevant? In what way could the p value be combined with other information? How could it be used inductively? Finally, how could one "reject" the null hypothesis with no alternative to accept (26)?

THE NEYMAN-PEARSON HYPOTHESIS TEST

In 1928, the mathematicians Jerzy Neyman and Egon Pearson published a landmark paper on the theoretical foundation for a procedure they called a "hypothesis test" (31). In one stroke, they seemed to solve many of the problems posed by Fisher's significance probability (26). They introduced the idea of the "alternative hypothesis" and its associated type II error. In a hypothesis test, one was to choose null and alternative hypotheses and the α and β error rates. These error rates were supposed to be tailored to a particular experimental situation, according to the consequences of each error (32), which contrasts with the almost universal use of $\alpha = 0.05$ today. These rates would define a "critical region" for the summary statistic (e.g., $Z > 1.96$). If a result fell into the critical region, then the alternative hypothesis was to be accepted and the null hypothesis rejected; if not, the null was to be accepted and the alternative rejected. This last characteristic contrasts with the oft repeated "one can never accept the null hypothesis, only fail to reject it," which was a feature not of the hypothesis test, but of Fisher's significance test, which had no alternative hypothesis.

There was no measure of evidence in the Neyman-Pearson hypothesis test, although some have attempted to interpret it that way (33). After an experiment, one was to report only whether the result fell in the critical region, not where it fell, as would be shown by a p value. This difference was not a minor one, for it represented a complete rejection of inductive reasoning. Neyman and Pearson were acutely aware of this, as they showed in the introduction to one of their original papers:

> . . . No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.
>
> But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong. Here, for example, would be such a "rule of behaviour"; to decide whether H of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$, reject H, if $x \leq x_0$, accept H. Such a rule tells us nothing as to whether in a particular case H is true . . . But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false (32, pp. 290–1).

This is a remarkable passage, remarkable because it presents so straightforwardly a statistical fact that is so contrary to current scientific practice. It states that if we want to use only "objective" probability, i.e., the probability of data under a given hypothesis, we cannot infer from a single experiment anything about the truth of the hypothesis.

But isn't the *purpose* of statistical methods to enable us to make inferences about hypotheses from individual experiments? Doesn't the p value tell us something about the null hypothesis? As we shall see, the p value is a strange kind of probability, very different from the error rate of hypothesis tests, and one must bring in nonprobabilistic concepts to tell us what it says about the null hypothesis.

Neyman and Pearson held that the best we can do with deductive probability theory is a rule for statistically dictated *behavior*, which they claimed would serve us well in the long run. Whether we *believe* a hypothesis we "accept" is not the issue; it is only necessary that we *act* as though it were true. Neyman saw this not just as a model for statistics, but for the scientific method.

> In the past, claims have been made frequently that statistical estimation involves some mental processes described as *inductive reasoning*. . . . in the ordinary procedure of statistical estimation, there is no phase corresponding to the description of "inductive reasoning." . . . all reasoning is deductive and leads to certain formulae and their properties. A new phase arrives when we decide to apply these formulae and to enjoy the consequences of their properties. This phase is marked by an act of will (not reasoning) and, therefore, if it is desired to use the adjective "inductive" . . . it should be used in connection with the noun "behavior" rather than "reasoning" (34, p. 210).

FISHER'S REACTION

Fisher clearly saw what was at stake here. The difference between his p value and hypothesis tests, which he derisively called "acceptance procedures" and "decision functions," was not merely mathematical; they represented different visions of science. The modern-day importance of such differences is seen in the innumerable debates where the term "unscientific" is an epithet, and credibility is accorded only to those allowed to wear the "scientific" mantle. In his last book (30), Fisher returned to this issue repeatedly, stating his position in strong language, as was his style:

> The concept that the scientific worker can regard himself as an inert item in a vast cooperative concern working according to accepted rules is encouraged by directing attention away from his duty to form correct scientific conclusions, to summarize them and to communicate them to his scientific colleagues, and by stressing his supposed duty mechanically to make a succession of automatic "decisions." . . . The idea that this responsibility can be delegated to a giant computer programmed with Decision Functions belongs to a phantasy of circles rather remote from scientific research. The view has, however, really been advanced (Neyman, 1938) that Inductive Reasoning does not exist, but only "Inductive Behaviour"! (30, pp. 104–5).

Perhaps most distressing to Fisher was that his position was virtually obscured by the incorporation of p values into hypothesis tests:

> On the whole the ideas (a) that a test of significance must be regarded as one of a series of similar tests applied to a succession of similar bodies of data, and (b) that the purpose of the test is to discriminate or "decide" between two or more hypotheses, have greatly obscured their understanding . . . (30, pp. 45–6).
>
> . . . The conclusions drawn from [significance] tests constitute the steps by which the research worker gains a better understanding of his experimental material. . . . More recently, indeed, a considerable body of doctrine has attempted to explain, or rather to reinterpret, these tests on the basis of quite a different model, mainly as a means to making decisions in an acceptance procedure. The differences between these two situations seem to the author many and wide, and I do not think it would have been possible had the authors of this reinterpretation had any real familiarity with work in the natural sciences, or consciousness of those features of an observational record which permit of an improved scientific understanding . . . (30, pp. 79–80).

It would appear that the division between the two camps could not be deeper, nor the distinctions more sharply put. Yet the two approaches somehow became intertwined. This occurred because the p value served as a bridge that permitted their confounding. To understand this, we need to explore in

more depth the quantitative meaning and interpretation of the p value.

p VALUE AS AN "OBSERVED ERROR RATE"

We will first examine the interpretation of the p value as an "observed error rate." That interpretation is reflected in textbooks and articles, some of which differ in their definitions of this fundamental measure:

> The p value is the observed α level . . . the smallest level of significance α at which an experimenter . . . would reject [the null hypothesis] on the basis of the observed outcome . . . (35, pp. 171–2).

> . . . statistical methods are used to define the probability that a false positive outcome (type I error) of a study is due to sampling variability or chance. By convention, a "significant" difference between therapeutic outcomes is usually accepted when a probability or alpha level is less than 5% ($p < 0.05$) (36).

> The p value is the likelihood of a study being positive when the null hypothesis is true; it is analogous to the false-positive rate . . . of a diagnostic test (37, p. 2459).

Each of these passages suggests that a p value is akin, albeit not identical, to a type I error rate; the last mistakenly defines it as identical to α. This linkage is misleading. The significance level, α, is the probability of a set of future outcomes, represented by the "tail area" of the null distribution. Implicit in the concept is that we don't know which of those outcomes will occur. The tail area represented by the p value is quite different; we know the outcome, and by definition it lies exactly on the border of the tail area. The "error rate" concept requires that a result can be anywhere within the tail area, which is not the situation with the p value. An error rate interpretation of the p value implies partial ignorance about the results, but if we had such ignorance, we could not calculate the p value.

A real-world example illustrates this. Suppose a school reports class rank in the following manner. Any student at the Nth percentile of the class is said to be "within the top N percent." For example, if a particular student ranks 15th out of 100 students, that student is reported as being "within the top 15 percent" of the class. The "15 percent" cited in this way falsely implies ignorance about the student's actual rank and suggests that the reporting threshold was set independently of the student. This same illusion is created when we interpret the p value as an "observed" error rate. Also like the p value, this "15 percent" has the peculiar property of including no one else in the class; everyone with a higher class rank will report being in a higher percentile.

Another factor that makes the error rate interpretation of p values problematic is that they are usually calculated conditionally. Fisher felt that aspects of the data irrelevant to the effect under study should not affect the p value. This is done by treating the irrelevant data (such as the marginal totals in contingency tables) as though they were fixed before the experiment and is implicit in many standard methods, such as linear regression and the Fisher exact test. However, conditioning on a quantity unknown until after the experiment means that the postexperiment p value cannot be a pure reflection of the preexperiment, unconditional alpha (24, 38, 39).

Confusion about what p values represent is reflected in the many styles of reporting them. One style is to state only whether p is less or greater than 0.05. This is the pure hypothesis test perspective, linking the imprecisely reported p value directly to the pretrial type I error rate. Another method, which is more informative about the observed data but breaks the link to type I error, is to quote the exact p value, e.g., "$p = 0.02$." In an effort to have it both ways, intermediate approaches have arisen. One is the "roving alpha" style, in which the investigator classifies the p value as falling into one of several fixed categories, usually $p < 0.05$, $p < 0.01$, and $p < 0.001$. Another involves expressing the exact p value as an inequality, i.e., writing "$p < 0.02$" when $p = 0.02$. None of these approaches are optimal, since we will see that the shift from the pre- to postexperiment perspective involves more

than juggling an inequality sign; it should mean changing the number itself.

If p values are not a form of error rate, what are they? We will now examine Fisher's proposal for p values as measures of inductive evidence in single experiments, or in his words, "a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test" (30, p. 47).

p VALUES AND LIKELIHOOD

To examine the inferential meaning of the p value, we need to review the concept of inductive evidence. An inductive measure assigns a number (a measure of support or credibility) to a hypothesis, given observed data. Inductive statistical evidence can be defined as the relative inductive support given to two hypotheses by the data (40). By this definition the p value is not an inductive measure of evidence, because it involves only one hypothesis and because it is based partially on unobserved data in the tail region (14, 20, 41).

To assess the quantitative impact of these philosophical issues, we need to turn to an inductive statistical measure: mathematical likelihood. This was the second inferential method promoted by Fisher. It is computationally simple, but conceptually subtle. We must first distinguish likelihood from its vernacular usage as a synonym for probability. When we speak of the "likelihood of a hypothesis," we are speaking of its support by the data, not its "probability of being true." The likelihood of a hypothesis, given the observed data, is proportional to the probability of the observed data, given that hypothesis (11, 14, 20, 42). Even though probability and likelihood are closely linked, Fisher made their distinction clear:

> Mathematical likelihood is not, of course, to be confused with mathematical probability... like mathematical probability, [likelihood] can serve in a well-defined sense as a "measure of rational belief"; but it is a quantity of a different kind than probability, and does not obey the laws of probability. Whereas such a phrase as "the probability of A or B" has a simple meaning... the phrase "the likelihood of A or B" is more parallel with "the income of Peter or Paul"—you cannot know what it is until you know which is meant.
>
> ... The likelihood supplies a natural order of preference among the possibilities under consideration ... (30, pp. 72–3).

The ratio of likelihoods (or their logarithm) can be used as a measure of the relative evidential support given by the data to two hypotheses (11, 40). The likelihood ratio is computed by taking the ratio of the data's probability under the null hypothesis to its probability under a specific alternative hypothesis. Likelihood ratios are used in Bayes' theorem (posterior odds = likelihood ratio × prior odds) and are a familiar part of the mathematics of screening and diagnostic tests (43). Even though the likelihood ratio is part of Bayes' theorem, it is completely separate from the prior odds of hypotheses. It is the part where "the data speak." Many epidemiologists are already familiar with this measure: the deviance, a general index of model fit, is twice the logarithm of a likelihood ratio of two nested models, and in genetic epidemiology, the Lod score is a log likelihood ratio (44).

We will use likelihood ratios to quantify the difference between two states of knowledge, represented by precise and imprecise p values (e.g., $p = 0.05$ vs. $p \leq 0.05$). For simplicity, we will assume that the p values are unconditional. In that case, when $p \leq 0.05$, it is correct to say that under the null hypothesis, an event with a 5 percent probability has occurred. However, when $p = 0.05$, many epidemiologists tend to make the same claim; it is hard to know what else to say. Because p values allow us to distinguish these situations only via the inequality sign, a quantitative difference in the evidence against the null hypothesis is not generally appreciated. Likelihood ratios will show that the difference is substantial.

In table 1, two likelihood ratios are compared, one for a "precise" p value, e.g., $p = 0.03$, and one for a corresponding hypothesis test, e.g., $p \leq \alpha = 0.03$. The alternative hypothesis used here is the one against which the hypothesis test has 90 percent power

TABLE 1. Ratio of the likelihood of the null hypothesis to the hypothesis of a difference associated with 90% power, $\alpha = 0.05$, under two different descriptions of the experimental result ($p = \alpha$, $p \leq \alpha$), shown for a range of p values, under a gaussian model

α	Z	Evidence for the null hypothesis vs. $\Delta_{0.05, 0.90}$		Minimum evidence for the null hypothesis when $p = \alpha$ (standardized likelihood)*
		$p = \alpha$	$p \leq \alpha$	
0.10	1.64	0.94	0.05	0.26
0.05	1.96	0.33	0.03	0.15
0.03	2.17	0.16	0.017	0.10
0.01	2.58	0.044	0.007	0.036
0.001	3.28	0.005	0.001	0.005

* The last column is the smallest possible likelihood ratio of the null hypothesis against any alternative.

(two-sided $\alpha = 0.05$), a typical choice in epidemiologic research. We assume that we know the direction of the effect. The likelihood ratio for the precise p value corresponds to the ratio of heights of the two probability densities at the observed data. The likelihood ratio for the imprecise p value is the ratio of areas of the two probability densities beyond the observed data (figure 1; see Appendix for mathematical details).

With this alternative hypothesis, "$p \leq 0.05$" represents *11 times* (= 0.33/0.03) less evidence in support of the null hypothesis than does "$p = 0.05$." Using Bayes' theorem, with initial probabilities of 50 percent on both hypotheses (i.e., initial odds = 1), this means that after observing $p = 0.05$, the probability that the null hypothesis is true falls only to *25 percent* (= 0.33/(1+ 0.33)). When $p \leq 0.05$, the truth probability of the null hypothesis drops to 3 percent (= 0.03/(1 + 0.03)). Below p values of 0.001, there is not much practical difference between the two situations, but in the critical range of 0.001–0.05, the differences are of a magnitude that could qualitatively affect the conclusions we draw from data. When we use the tail region to represent a result that is actually on its border, we misrepresent the evidence, making the case against the null hypothesis look much stronger than it actually is.

Since Fisher simultaneously promoted p values and likelihood, one might have expected him to make clear the advantage of the p value definition, but he never did. As the philosopher Ian Hacking noted, "At no time does Fisher state why one is allowed to add the clause 'or a greater value' so as to form the region of rejection" (20, p. 82). Some insight into his thinking might be gained from the following comment on confidence intervals (italics added; "p" refers to the probability parameter of a binomial distribution, not the p value):

> Objection has sometimes been made that the method of calculating confidence limits by setting an assigned value such as 1% on the frequency of observing 3 or less . . . is unrealistic in treating values less than 3, which have not been observed, in exactly the same manner as 3, which is the one that has been observed. *This feature is indeed not very defensible save as an approximation.* It should be pointed out that when the probability of 3 or less is small, most of this small probability will be due to the case "exactly 3," and that the contribution of the other three cases is not very important, although it does increase or decrease with varying p at a relative rate different from the contribution of "exactly 3" itself . . . It would, however, have been better to have compared the different possible values of p, in relation to the frequencies with which the actual values observed would have been produced by them, as is done by Mathematical Likelihood . . . (30, p. 71).

As with many things Fisher wrote, the preceding passage can be interpreted in a variety of ways. Since one can show a relation between p values and the tail area discussed above, one could argue that Fisher was implicitly acknowledging here that the

p value was an "approximate" attempt, without an alternative hypothesis, to get information like that provided by likelihood ratios. This is consistent with his vagueness about the quantitative interpretation of the p value, his stress on its informal use, and his fury and frustration at seeing it subsumed by hypothesis testing.

THE STANDARDIZED LIKELIHOOD

Can likelihood ratios be used in lieu of p values? Since every alternative hypothesis has a different likelihood, the comparison of each of those alternative hypotheses to the null hypothesis yields a different likelihood ratio. One proposal, promoted by Fisher and others (11, 19, 45), is to use the likelihood ratio of the null hypothesis to the unique hypothesis with maximum likelihood: with gaussian data, the hypothesis that the true population values are equal to the observed estimates. This ratio is called the standardized likelihood. In figure 1, this would mean using the alternative hypothesis whose probability density is centered on the observed mean. Since this alternative has the highest likelihood, the standardized likelihood represents the smallest amount of statistical evidence that can be attributed to the null relative to any alternative. It is a "worse case scenario" for the null. It has a Bayesian interpretation as the smallest factor, after seeing the data, by which one can multiply the prior odds of the null hypothesis to get the final odds.

For gaussian data, the standardized likelihood is equal to $\exp(-Z^2/2)$, where Z is the number of standard errors from the null hypothesis. In table 1, we see that the weakest evidence for the null hypothesis is still 3–5 times higher than the associated 2-sided p value. (Using one-sided p values would double the disparity.) When $p = 0.05$, the support that can be mustered for the best alternative is only 6.7 times (1/0.15) the support for the null hypothesis. This means that, if the null hypothesis has initial odds of 1.0, $p = 0.05$ makes the final odds no lower than 0.15, corresponding to a probability of $1/(1 + 0.15) = 0.13$. In order for the final probability of the null hypothesis to be 5 percent (final odds = 1/19) after observing $p = 0.05$, its initial probability can be no higher than 26 percent. If a relation is thought to be improbable, corresponding perhaps to an initial null probability of 80 percent, a $p = 0.05$ would lower this probability only to 38 percent. The standardized likelihood needed to make it only 5 percent probable is $5/95 \div 80/20 = 1/76 = 0.013$, corresponding to a p value of 0.003. So even the strongest quantitative case against the null is not nearly as strong as the p value would indicate (18, 19, 46).

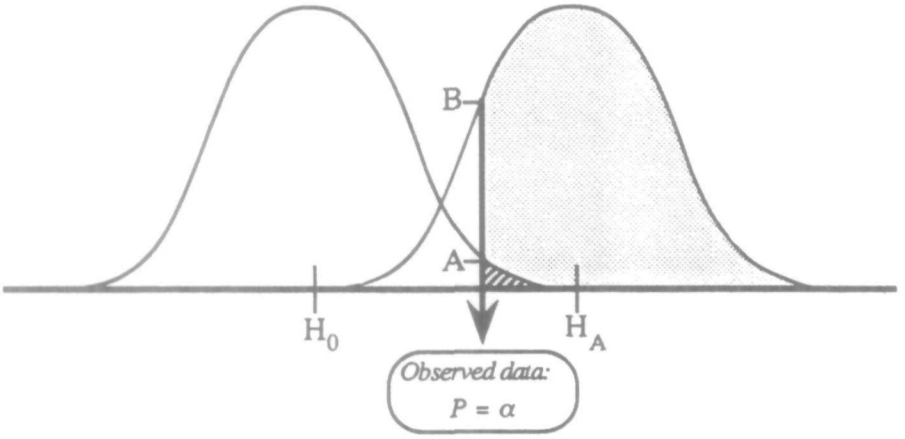

FIGURE 1. Graphical representation of the derivation of table 1. The curves are the gaussian probability densities for outcomes under H_0 and H_A, the null and alternative hypotheses. The likelihood ratio associated with the precise p value ($p = \alpha$) is A/B, the ratio of the curve heights at the observed data. The likelihood ratio associated with the imprecise p value ($p \leq \alpha$) is the ratio of the small striped area to the total shaded area (including the striped area).

The standardized likelihood seems to more accurately represent the informal weight that is put on p values. Epidemiologists usually describe p values in the 0.02–0.05 range as representing only moderate evidence against the null. That description is better reflected by the corresponding range of the gaussian standardized likelihood, 0.07–0.15, than by 0.02–0.05. An implausible association with a p value in that range will often not be seriously considered, consistent with the calculations above. It has been suggested that Fisher himself tended to use in practice a rejection threshold of $p = 0.01$, which corresponds to a standardized likelihood of 0.04, in keeping with the notion that a "1 in 20" ratio should be the evidential threshold to reject the null hypothesis (28). Because the standardized likelihood usually varies monotonically with the p value, the issue of the p value's three- to fivefold overstatement of the evidence is sometimes dismissed as a problem of "calibration." Imagine explaining that to a new student, to a policymaker, or in a court of law.

IMPLICATIONS FOR SCIENCE

We have seen that there are serious problems with both the error rate and evidential interpretations of the p value. When we combine them, insisting that the p value must reflect the "correct" type I error to properly represent the evidence, we create a potent illusion that produces a host of paradoxes and problems, although their source is rarely recognized. They include the "multiple looks" problem (47–50), the multiple comparisons problem (51), how sample size affects the interpretation of a given p value (25, 52), the probability of replicating a significant finding as a function of the p value (53), whether one-sided p values should be used for one-sided hypotheses (54), and the appropriate thresholds for meta-analyses (55).

An important aspect of this analysis is the insight it gives into Fisher and Neyman's primary concern and source of conflict: the nature of the scientific method. Neyman's position was that we should set up rules with pretrial error rates for hypothesis rejection and "enjoy the consequences" of their use. The p value's continuous scale undercut this by inviting the use of informal induction regardless of where the pretrial rejection threshold was set. A p value of 0.04 produces a different reaction than one of 0.00001, though both are significant at $\alpha = 0.05$. The nature of that different reaction is outside the domain of deductive probability theory, and therefore, according to Neyman, outside the realm of the objective scientific method.

Also inimical to Neyman's position was the error rate interpretation of the p value. It implied that, if $p = 0.001$, one could state that one was "enjoying the consequences" of using a rule with $\alpha = 0.001$. This undermined the reason that error rates had to be set before the experiment, vitiating the force of Neyman's rules for the objective conduct and interpretation of research. The obligatory statement in most research articles, "p values below 0.05 were considered statistically significant," is an empty exercise in semantics. It tells us nothing, only that the word "significant" will be used to refer to relations where $p \leq 0.05$. If only "statistical significance" was reported, without p values, then the declaration of the pretrial α would be critical.

The interpretation of the p value as an observed α was even more damaging to Fisher's position than it was to Neyman's. It facilitated the incorporation of the p value into the hypothesis test framework, which we have seen was anathema to Fisher. It also implied that evidence and uncertainty about hypotheses could be described in the language of unconditional, pretrial probability. As Fisher wrote in a letter to a colleague, "... the concept of mathematical probability is inadequate to express the nature and extent of our uncertainty in the face of certain types of observational material, while in all cases the concept of mathematical likelihood will supply very helpful guidance, if we are prepared to give up our irrational urge to express ourselves only in terms of mathematical probability" (56, p. 92).

This "irrational urge" (or the lack of alter-

natives) may be the source of the dilemma many epidemiologists face when attempting to express uncertainty in policy settings. Suppose we perform a study and obtain $p = 0.03$ for an unlikely relation. Summarizing this as "if there is no relationship, there is a 3 percent probability that the observed or larger effect could have been obtained due to chance" may not do justice to either our informal assessment of the meaning of $p = 0.03$ or our uncertainty about the relation. It also invites the misinterpretation that there is a 97 percent chance that the effect is real. Noting that the gaussian standardized likelihood for $p = 0.03$ is 0.10, it would be more appropriate to say something like, "the plausibility of some relation, relative to the plausibility of no relation, is at most tenfold greater than it was before the experiment." This uses a correct quantitative measure of evidence, calls attention to its comparative nature, and highlights the importance of the prior plausibility of the association.

CONCLUSION

The originators of the statistical frameworks that underlie modern epidemiologic studies recognized that their methods could not be interpreted properly without an understanding of their philosophical underpinnings. Neyman held that inductive reasoning was an illusion and that the only meaningful parameters of importance in an experiment were constraints on the number of statistical "errors" we would make, defined before an experiment. Fisher rejected mechanistic approaches to inference, believing in a more flexible, inductive approach to science. One of Fisher's developments, mathematical likelihood, fit into such an approach.

The p value, which Fisher wanted used in a similar manner, invited misinterpretation because it occupied a peculiar middle ground. Because of its resemblance to the pretrial α error, it was absorbed into the hypothesis test framework. This created two illusions: that an "error rate" could be measured after an experiment and that this posttrial "error rate" could be regarded as a measure of inductive evidence. Even though Fisher, Neyman, and many others have recognized these as fallacies, their perpetuation has been encouraged by the manner in which we use the p value today. One consequence is that we overestimate the evidence for associations, particularly with p values in the range of 0.001–0.05, creating misleading impressions of their plausibility. Another result is that we minimize the importance of judgment in inference, because its role is unclear when postexperiment evidential strength is thought to be measurable with preexperiment "error-rates." Many experienced epidemiologists have tried to correct these problems by offering guidelines about how p values should be *used*. We may be more effective if, in the spirts of Fisher and Neyman, we instead focus on clarifying what p values *mean*, and on what we mean by the "scientific method."

REFERENCES

1. Weed DL. On the logic of causal inference. Am J Epidemiol 1986;123:965–79.
2. Susser M. The logic of Sir Karl Popper and the practice of epidemiology. Am J Epidemiol 1986; 124:711–18.
3. Rothman K. Causal inference. Chestnut Hill, MA: Epidemiology Resources, 1988.
4. Rothman K. Modern epidemiology. Boston, MA: Little, Brown, 1988.
5. Rothman K. Significance questing. Ann Intern Med 1986;105:445–7.
6. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989; 79:340–9.
7. Cutler S, Greenhouse S, Cornfield J, et al. The role of hypothesis testing in clinical trials. J Chronic Dis 1966;19:857–82.
8. Walker AM. Reporting the results of epidemiologic studies. Am J Public Health 1986;76:556–8.
9. Anscombe F. Sequential medical trials. J Am Stat Assoc 1963;58:365–83.
10. Pratt J. Bayesian interpretation of standard inference statements. J R Stat Soc B 1965;27:169–203.
11. Edwards A. Likelihood. Cambridge: Cambridge University Press, 1972.
12. Savage L. The foundations of statistical inference: a discussion. New York: Wiley, 1962.
13. Barnard G. The Bayesian controversy in statistical inference. J Inst Actuaries 1967;93:229–69.
14. Goodman S, Royall R. Evidence and scientific research. Am J Public Health 1988;78:1568–74.
15. Poole C. Beyond the confidence interval. Am J Public Health 1987;77:195–9.

16. Berger J, Berry D. Statistical analysis and the illusion of objectivity. Am Scientist 1988;76:159–65.
17. Diamond G, Forrester J. Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med 1983;98:385–94.
18. Berger J, Sellke T. Testing a point null hypothesis: the irreconcilability of p values and evidence. J Am Stat Assoc 1987;82:112–39.
19. Berger J. Are p values reasonable measures of accuracy? In: Francis IS, Manly BFJ, Lam FC, eds. Vol I. Amsterdam: North-Holland/Elsevier, 1986.
20. Hacking I. The logic of statistical inference. Cambridge: Cambridge University Press, 1965.
21. Cox D, Hinckley D. Theoretical statistics. Cambridge: Chapman and Hall, 1974.
22. Kyburg H. The logical foundations of statistical inference. Dordrecht, Holland: D. Reidel, 1974.
23. Cox D. The role of significance tests. Scand J Stat 1977;4:49–70.
24. Seidenfeld T. Philosophical problems of statistical inference. Dordrecht, Holland: D. Reidel, 1979.
25. Oakes M. Statistical inference: a commentary for the social sciences. New York: Wiley, 1986.
26. Howson C, Urbach P. Scientific reasoning: the Bayesian approach. La Salle, IL: Open Court, 1989.
27. Johnstone D. Tests of significance in theory and practice. Statistician 1986;35:491–504.
28. Salsburg D. Hypothesis versus significance testing for controlled clinical trials: a dialogue. Stat Med 1990;9:201–11.
29. Box J. RA Fisher: the life of a scientist. New York: Wiley, 1978.
30. Fisher R. Statistical methods and scientific inference. 3rd ed. New York: Macmillan, 1973.
31. Neyman J, Pearson E. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 1928;20:175–240.
32. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A 1933;231:289–337.
33. Birnbaum A. The Neyman-Pearson theory as decision theory and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. Synthese 1977;36:19–49.
34. Neyman J. Lectures and conferences on mathematical statistics and probability. 2nd ed. Washington, DC: US Department of Agriculture, 1952.
35. Bickel P, Docksum K. Mathematical statistics. San Francisco: Holden-Day, 1977.
36. Barnes R. Who took the "p" out of statistics? J Vasc Surg 1989;10:100–3.
37. Browner W, Newman T. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987;257:2459–63.
38. Berger J. The frequentist viewpoint and conditioning. In: LeCam L, Olshen R, eds. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer. Vol 1. Belmont, CA: Wadsworth, 1985:15–43.
39. Greenland S. On the logical justification of conditional tests for two-by-two contingency tables. Am Stat 1991;45:248–51.
40. Good I. Probability and the weighing of evidence. New York: Charles Griffin & Co., 1950.
41. Berkson J. Tests of significance considered as evidence. J Am Stat Assoc 1942;37:325–35.
42. Birnbaum A. On the foundations of statistical inference (with discussion). J Am Stat Assoc 1962; 57:269–326.
43. Sackett D, Haynes R, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown, 1985.
44. McCullagh P, Nelder J. Generalized linear models. 1st ed. New York: Chapman and Hall, 1983.
45. Barnard G. The use of the likelihood function in statistical practice. In: Proceedings of the V Berkeley Symposium. Vol 1. Berkeley: University of California Press, 1966:27–40.
46. Edwards W, Lindman H, Savage L. Bayesian statistical inference for psychological research. Psychol Rev 1963;70:193–242.
47. Dupont W. Sequential stopping rules and sequentially adjusted p values: Does one require the other? (with discussion). Controlled Clin Trials 1983;4: 3–10.
48. Lindley D. A statistical paradox. Biometrika 1957; 44:187–212.
49. Cornfield J. A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. J Am Stat Assoc 1966;61:577–94.
50. Cornfield J. Sequential trials, sequential analysis, and the likelihood principle. Am Statistician 1966; 20:18–23.
51. Thomas DC, Siemiatycki J, Dewar R, et al. The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol 1985; 122:1080–95.
52. Royall R. The effect of sample size on the meaning of significance tests. Am Statistician 1986;40: 313–15.
53. Goodman S. A comment on replication, p values, and evidence. Stat Med 1992;11:875–9.
54. Goodman S. One or two-sided p values? Controlled Clin Trials 1988;9:387–8.
55. Goodman S. Meta-analysis and evidence. Controlled Clin Trials 1989;10:188–204,435.
56. Bennett JH. Statistical inference: selected correspondence of RA Fisher. New York: Wiley, 1991.

(Appendix follows)

APPENDIX

The likelihood ratios that appeared in table 1 are calculated as follows:

$$\text{Testing } H_0\text{: } \mu = 0 \text{ vs. } H_A\text{: } \mu = \Delta_{0.05,0.90}$$

where $\Delta_{0.05,0.90}$ is the difference against which the hypothesis test has two sided $\alpha = 0.05$ and one-sided $\beta = 0.10$ (power = 0.90). The Z score corresponding to this alternative hypothesis, Z_Δ, equals 1.96 + 1.28 = 3.24. Because we are comparing a precise p value to a corresponding hypothesis test, the observed Z score will be designated by Z_α.

The likelihood ratio (LR) for the imprecise p value, corresponding to the ratio of shaded areas in figure 1, is

$$\text{LR}(H_0 \text{ vs. } H_A \mid p \leq \alpha) = \frac{\alpha/2}{1 - \Phi(Z_\alpha - 3.24)}$$

where $\Phi(Z)$ is the area under the gaussian curve to the left of Z.

The likelihood ratio for the precise p value, corresponding to A/B in figure 1, is

$$\text{LR}(H_0 \text{ vs. } H_A \mid p = \alpha) = \frac{ke^{-Z_\alpha^2/2}}{ke^{-(3.24-Z_\alpha)^2/2}} = e^{5.25 - 3.24Z_\alpha}$$