Commentary

Evidence and Scientific Research

STEVEN N. GOODMAN, MD, MHS, AND RICHARD ROYALL, PHD

Abstract: This commentary reviews the arguments for and against the use of p-values put forward in the *Journal* and other forums, and shows that they are all missing both a measure and concept of "evidence." The mathematics and logic of evidential theory are presented, with the log-likelihood ratio used as the measure of evidence. The profoundly different philosophy behind evidential methods (as compared to traditional ones) is presented, as well as a comparative example showing the difference between the two approaches. The reasons why we mistakenly ascribe evidential meaning to p-values and related measures are discussed. Unfamiliarity with the technology and philosophy of evidence is seen as the main reason why certain arguments about p-values persist, and why they are frequently contradictory and confusing. (*Am J Public Health* 1988; 78:1568–1574.)

Introduction

The series of articles[1–4] in the *Journal* on p-values and confidence intervals has been a useful introduction to a debate that has been conducted mainly in the philosophical and statistical literature for the last 50 years. The main focus of the *Journal* discussion was the appropriate roles and uses for p-values, confidence intervals, and other statistical measures in summarizing the results of epidemiologic studies. Where Fleiss generally defended the use of p-values,[1] Walker,[2] Thompson,[3] and Poole[4] criticized them.

Both Walker and Thompson urged the use of confidence intervals (CI) instead of isolated p-values because they convey more information, i.e., the actual magnitude of the measured effect as well as the precision of the estimate. Poole pointed out that the CI is often used in a way that differs little from the significance test—i.e., we simply look and see whether or not the null value is included in the interval. He discussed the needs of decision makers, and encouraged the use of complete p-value curves as the most complete display of what the data say while also freeing us from seemingly arbitrary statistical dictates. All three of these authors shared an aversion to rigid rules of interpretation that bypass judgment (like the $p < .05$ threshold) or offend scientific intuition (like adjustments for multiple comparisons).

It appears that what motivated those writers was a feeling that something important about the data is not being captured either by the p-value, or by the manner in which it is used. All of the discussions, including Fleiss's, reflected a desire to have an index that objectively summarizes data and that helps the scientist or decision maker interpret patterns. It was obvious to all that the p-value is an imperfect statistical summary index, but except for a hint in Poole's essay, there was not a discussion of alternatives. The commentators decried the uncomfortable shackles of the p-value concept, but they were all forced to return to it in one form or another. In the end, the issue was not what the p-value *means*, but how it should be *used*.

Is there an alternative to the p-value? The answer is yes. A substantial body of research over the last 40 years has gone into the definition and measurement of the "weight of evidence", which scientists can use to make decisions. But this "weight of evidence" involves a very different concept of the relationship between theory and observations than the p-value represents. In this paper we will describe this concept, show how it exposes the logical and inferential problems inherent in p-values, and produces results in closer accord with our scientific intuition. We believe that the evidential concept is what Walker, Thompson, and Poole were striving for, and that it offers a theoretical framework upon which their recommendations for the use of p-values are based, and upon which many others give the advice that p-values should not be used at all.

Evidence is a property of data that makes us alter our beliefs about how the world around us is working. Another way to say this is that evidence is the basis upon which we derive inferences. Can the p-value provide such a basis? To answer that question, we need to explore the central logical tenet that justifies its use: that if an observation is rare under a hypothesis then it can be regarded as evidence against that hypothesis.

Consider the following examples. Suppose one draws a queen of spades and a seven of clubs from a well-shuffled deck of cards. Under the hypothesis that the deck is a normal one, this event is rare, with a probability of only .0004. Is this evidence that the deck is not normal? Now suppose you're in a casino and see the numbers 3, 14, 6 and 27 come up consecutively on a roulette wheel. Under the hypothesis that the wheel is fair, the probability is only $(1/38)^4 = .0000005$; should we interpret this as evidence against that hypothesis? Finally, what if a previously unsuspected association shows up in a study with a $p = .01$? Do we rush to publish? (Unfortunately, while the answer to the first two questions is probably "no", the last one may be "yes".)

These scenarios show that we do not automatically interpret events that are rare under a specific hypothesis as evidence against that hypothesis. Life is full of rare events to which we accord scant attention. What *does* make us react is a plausible competing hypothesis under which the data are more probable. Suppose you find out that the person who handed you the cards owns a trick deck with only sevens and

From the Departments of Epidemiology and Biostatistics, Johns Hopkins School of Hygiene and Public Health. Address reprint requests to Steven N. Goodman, MD, Department of Epidemiology, School of Hygiene and Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205. This paper, submitted to the Journal February 16, 1988, was revised and accepted for publication July 18, 1988.

Editor's Note: See also related editorial p 1531 this issue.

queens. Suddenly, your draw becomes strong evidence in favor of the trick deck alternative vs the normal deck hypothesis. In the casino, you note that the four numbers on the wheel are adjacent, suggesting another hypothesis that makes the observed sequence more probable—that the wheel is weighted or biased in some way. Finally, suppose a reviewer proposes a plausible biological explanation for the association you measured. It is only the existence of this alternative that elevates the data to the status of reportable "evidence" against chance and, for biology, being at work.

Even though we often use the p-value as part of a hypothesis-testing procedure in which we can "reject the null" and "accept the alternative", the p-value itself has no information about that alternative; it is defined as the probability of what we saw, plus "more extreme" results, *only under the null hypothesis*. That the p-value fails to represent evidence because it depends only on one hypothesis is hardly a new idea in statistical circles. It has been around for over 50 years, as reflected in the quote below from Gossett (the inventor of the t-test) from the 1930s:

> . . .[a significance test] doesn't in itself necessarily prove that the sample is not drawn randomly from the population even if the . . . [p-value] is very small, say .00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 . . . you will be very much more inclined to consider that the original hypothesis is not true. (in Hacking[5], p 83)

The p-value is not adequate for inference because the measurement of evidence requires at least three components: the observations, and two competing explanations for how they were produced. In scientific research, these competing explanations usually take the form of the null and alternative statistical hypotheses. No matter how rare our data are under the null hypothesis, they remain mere numbers, not yet inferential "evidence", until we can propose another hypothesis that also explains them.

Philosophic Issues

The issue posed above relates to a fundamental philosophic question; can non-relative, purely negative evidence, which the p-value represents, play a role in scientific evaluation of theories? The philosopher Karl Popper would say yes. His view is that science moves forward by successive disproofs that do not require competing theories.[6] The 20th century philosophers Karl Hempel and Rudolf Carnap felt that formal induction was an essential part of the scientific process.[7,8] Induction carries with it a concept of positive, relative evidence, whereby a theory is not just "falsifiable", but might be made *more* likely (relative to another theory) by observations. The philosophical debate is complex, and its details are beyond the scope of this paper. But it is important to recognize that a statistical index (representing absolute or relative evidence) can carry with it a specific view of the scientific process. Our contention is that the conflict between the inductivist view, which many scientists informally hold, and the one embodied by the p-value usually goes unrecognized and is the source of much confusion.

Consequences of the P-value Definition

The most widely recognized practical consequence of the p-value's dependence on only one hypothesis is that a huge effect in a small trial or a minuscule effect in a large trial can result in identical p-values. To the extent that we believe the size of an effect is an essential part of the evidence relative to the hypothesis of "no effect", then the p-value is inadequate for measuring the strength of evidence.[9] The move toward confidence intervals is an effort to deal with this issue by focusing on the effect size.

Another problem with using the p-value to measure evidence is caused by its inclusion of "more extreme values" in its calculation. With a p = .03 we are taught to say, "If the null hypothesis were true, and we repeated this experiment many times, we would observe a difference this big or bigger 3 per cent of the time." But the "bigger" values in the tail of a distribution are not just more extreme, they are *unobserved*. In other words, the rarity of what we did see is assessed by combining it with the probability of results that didn't happen—in Jeffreys' words, "A hypothesis that may be true may be rejected because it has not predicted . . . results which have not occurred."[10]

Aside from being a logical conundrum, the above issue results directly in a practical problem, because what is "more extreme" depends on how an experiment was conducted. Suppose we consider two trials: a fixed sample-size trial comparing two treatments on each of 10 people, and another one using the same treatments and the same 10 people, except that this time the plan is to stop as soon as treatment A is better than treatment B in seven of the patients. Now suppose the seventh patient showing a preference for A is the tenth patient in both trials. In the fixed-size trial the "more extreme" results are those with eight or more preferences for A. In the other trial, the "more extreme" area consists of situations where it takes less than 10 patients to reach seven preferences for A. These regions can have quite different probabilities, so these experiments will have different p-values *even though they yielded identical results on the identical subjects*.

The only difference between the two trials above was what the experimenter would have done if the observations had been different; would s/he have stopped the trial after only seven patients if all seven had shown a preference for A? If the scientist died without telling anyone what s/he would have done, traditional teaching would tell us that we could not calculate a valid p-value. In studies where p-values (or CIs) are used we must somehow learn the investigator's possibly hidden actions and intentions during the trial, so "what the data say" is often obscured by questionable answers to unanswerable questions.[9]

The above difficulties are not just encountered when conducting trials. The most familiar example of the p-value dependence on our state of mind is in the choice of one vs two-sided statistical tests. Even though two-sided tests are used most of the time, students are taught that if they would not consider a difference in one direction a possibility, they can legitimately use a one-sided p-value, which is half the size. Again, the data are the same, but the p-value is changed by a profoundly subjective consideration. A similar situation arises when multiple comparisons are made. In summary, the p-values can be equal in situations where we feel the evidence is very different (different effect magnitude in trials with different sample sizes), or different in situations where we would think the evidence should be the same (same data in trials with different stopping rules).

The Likelihood Axiom

We will now look at an alternative basis for inference, the likelihood function. It is a measure that has few of the difficulties inherent in the p-value, and unlike the p-value, has

a sound theoretical foundation as a basis for inference.[11,12] It is defined as follows, with "c" designating an arbitrary constant, characteristic of the statistical model being used:

$$\text{Likelihood } (H \mid \text{Data}) = \underset{\text{Discrete case}}{c*\text{Prob}(\text{Data} \mid H)} = \underset{\text{Continuous case}}{c*f(\text{Data} \mid H)}$$

In the notation above, H represents a specific hypothesis, and "f(Data | H)" is the probability density function of the data under H. The entire likelihood function is obtained by calculating the likelihood over the range of hypotheses, which are usually defined in terms of some average characteristic of a population. At first glance, the likelihood may look simply like a relabeled probability. But a likelihood is profoundly different from a probability in that the data are regarded as fixed and it is the hypotheses that are variable, whereas probabilities are calculated assuming a fixed hypothesis and random data. Likelihoods also do not obey the laws of probability. Finally, as signified by the arbitrary constant, an isolated likelihood does not have a unique value—the scale of the likelihood function is arbitrary.

We need a prescription for how to use and interpret the likelihood function. This is found in the "Likelihood Axiom", stated by Edwards:

> Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other.[13]

This axiom tells us that the evidence supporting one hypothesis versus another is represented by the ratio of their likelihoods. The likelihood's arbitrary constant means that we cannot speak of the absolute support (or lack thereof) of a single hypothesis by any set of data, and so we cannot use the data to "reject" it. The *relative* support is uniquely defined however, since the constant cancels out when we take a ratio.

The likelihood ratio (LR) has a close relationship with fundamental measures in other sciences (information in communications theory, entropy in physics, odds ratio in epidemiology) and its rediscovery by Alan Turing in the 1940s was the key to his breaking the German Enigma code in WWII.[14] Having already defined likelihood, the mathematical definition of the LR is straightforward:

$$\text{Likelihood ratio (For } H_0 \text{ vs } H_1 \mid \text{Data}) = \frac{c*\text{Prob}(\text{Data} \mid H_0)}{c*\text{Prob}(\text{Data} \mid H_1)} = \frac{c*f(\text{Data} \mid H_0)}{c*f(\text{Data} \mid H_1)}$$

Discrete case — Continuous case

The LR is equivalent to the ratio of the data's probability under one hypothesis compared to its probability under another hypothesis. Graphically, it is simply the ratio of heights of the two hypothesized probability distributions at the observed data point (Figure 1). Thus we can restate the likelihood axiom in plain language: The hypothesis better supported by the data is the hypothesis which better predicts the data.

Suppose two geneticists make different claims about the inheritance pattern of a rare disease in a family. One claims that the disease is transmitted in a fashion corresponding to a 75 per cent chance of transmission, whereas the other claims a 25 per cent chance. The parents then have a child who has the disease. What is unknown now is not whether the child will have the disease, but which hypothesis was the better one. We compare the two predictions through the LR, which equals 1/3 (=.25/.75), a result in favor of the 75 per cent prediction by a factor of three. The way we express this is that the observation (disease present) is evidence supporting the hypothesis of a 75 per cent chance of disease (with its underlying biologic explanation) in favor of a 25 per cent chance.

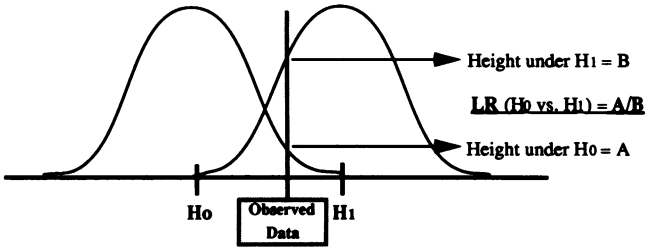

FIGURE 1—A Graphical Representation of the Calculation of the Likelihood Ratio (LR) for Two Simple Hypotheses Given Experimental Data under a Single Statistical Model
Even though these curves represent probability distributions, the likelihoods are defined only at the observed data point, and this has arbitrary scale, hence the y axis has no units. The one-sided p-value corresponding to these data would be the proportion of the area under the H_0 curve to the right of the data point. Different stopping rules can affect the shape of these curves and hence change that area, but the ratio of heights at the data point will remain the same.

The likelihood function and ratio is not affected by the reasons for stopping an experiment, nor by the number of other comparisons. Finally, and perhaps most important, the likelihood ratio has the same meaning in trials of different designs and sizes.

Evidence vs. Belief

If we say one hypothesis has a higher likelihood than another, we only mean that it is better supported by the evidence, **not** that it is more "likely" (i.e., probable) to be true. This distinction between what the evidence says and what we believe is a subtle but critical one. In the previous example, the hypothesis that the disease had a *100 per cent* chance of transmission is best supported by the fact of the child having the disease. This does not mean that we should believe the 100 per cent hypothesis, since it may be quite implausible biologically. This issue also arises in the labels assigned to "borderline" p-values (e.g., between .05 and .15), which often correspond to LRs in the weak-moderate range. Although to the author the data may represent a "trend", to some readers it may be a non-association. These labels reflect belief, not evidence. The frequent disputes about them could probably be minimized by saying, "The statistical evidence is moderate (or, the LR = ...), and we believe it represents a real phenomena because. . .", with a biologic or epidemiologic discussion to follow.

Another way to understand the distinction between belief and evidence is to look at the role the LR plays in Bayes Theorem, which can be written as follows:

$$\text{Final Odds} = \text{Initial odds} \times \text{Likelihood ratio}$$

Or, taking logs:

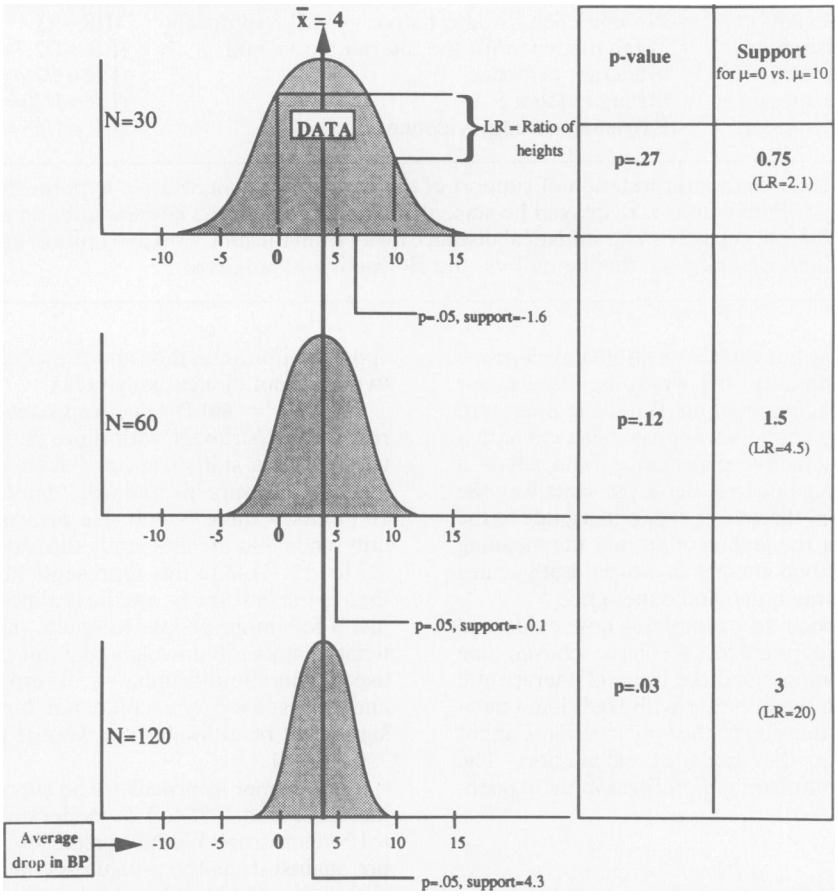

FIGURE 2—P-Value and Likelihood Measures of Evidence Provided by the same Observed Difference in Experiments of Three Different Sizes
The curves are the likelihood functions of the average fall in blood pressure from a drug given an observed four-point average drop. The LR for any pair of hypotheses is the ratio of curve heights at those hypotheses (it is calculated here for $\mu = 0$ vs $\mu = 10$). Positive support is evidence for no effect, a negative one is evidence for a ten point effect. Support corresponding to the p = .05 point is also marked on each curve. See text and Appendix for more details.

Final log odds (Final Beliefs) = Initial log odds (Initial Beliefs) + Log Likelihood Ratio (Weight of Evidence, Support)

Bayes Theorem is a mathematical identity; it describes how probabilities are changed by new statistical evidence. When probabilities are based on the frequency of random events, such as the chance of illness in a member of a population with a known disease prevalence, Bayes Theorem is indisputable. When the probabilities are measures of belief in a statistical hypothesis, the application of Bayes Theorem is called "Bayesian Inference" and is more controversial. However, the controversy centers on the appropriateness of probability as a measure of belief, not on the representation of evidence. The LR is the only way in which data enter the equation, and is clearly separated from the subjective factors that affect our initial odds. Nothing resembling a p-value is involved.

Bayes Theorem is a very useful tool to compare the quantitative meaning of LRs and p-values when the underlying probabilities are known. In such situations, when we compare p-values to Bayesian calculations on the same set of data, strong arguments have been made that the p-value almost always overstates of the degree of conflict with the null hypothesis.[15,16]

Bayes Theorem lies at the heart of quantitative decision analysis, making the LR the way to represent data as evidence in such analyses. Even when the decision maker does not want to use a formal decision calculus, we would submit that the LR or likelihood function remains the most appropriate data summary. This is in accord with Poole's point that the p-value is inappropriate for decision making.[4]

Support and Weight of Evidence

Before using the LR in a statistical problem, we need to have some way of interpreting the numerical results. A commonly used scale is the logarithm of the LR,[17,18] because this converts the likelihood ratio into a difference, and makes evidence additive, just as we naturally think of it. The log LR is called the "support", a word we will henceforth use in that technical sense.[13] A rough guide to support interpretation is shown at the top of the next page, which is partly based on the Bayesian calculations referred to earlier.

The above scale and some of the previous discussion of the meaning of the LR in real situations may sound uncomfortably vague. It might be helpful to see it as similar to temperature measurement. Few would argue that a thermometer does not measure thermal energy objectively. However, we all know that it is only a crude index of the subjective

0 units support	= No evidence for alternative vs null hypothesis	(LR=1)
-1 unit support	= **Weak** evidence for the alternative vs null	(LR=1/2.7=0.37)
-2 units support	= **Moderate** evidence	(LR=1/7.4=0.14)
-3 units support	= **Strong** evidence	(LR=1/20=0.05)
-4 units support	= **Extremely strong** evidence	(LR=1/55=0.02)

Informal guide to interpretation of support of the null over the alternative hypothesis. The range of negative units is displayed because this corresponds to the familiar situation when we are measuring increasing statistical distance away from the null. Positive units of support would indicate evidence for the null vs. the alternative hypothesis.

experience of "heat". How hot we feel on an 80° day depends on such factors as the humidity, the wind, the clouds, our ability to sweat, and our acclimatization. But the temperature still remains an invaluable objective, reproducible guide to a critical aspect of that subjective experience. The LR is a measure of evidence and a guide to belief in the same way the temperature is a measure of thermal energy and a guide to the sensation of heat. Neither the feeling of 80° nor the meaning of LR = 10 can be described exactly in words; both scales acquire their meaning through use and experience.[13]

We will now go through an example of how evidential measures can be used in practice. We have chosen one involving multiple experiments and the issue of therapeutic equivalence to emphasize the contrast with traditional measures. We hope to draw attention to the ways we *think* about these results, as well as the mode of calculation. The formulae underlying the numbers are outlined in the Appendix.

Example: Blood Pressure Regulation

Let us assume that a new drug is developed that can lower systolic blood pressure. Figure 2 shows the likelihood curves of three trials of different sample sizes, all of which show that the drug lowers the pressure by four points more than the standard therapy. Here are three scenarios using different analytic methods.

I. (N = 30) P-value Approach—The p = .27 is non-significant, so investigator A publishes his paper which "fails to reject" the hypothesis that the two drugs have equal effect. Scientist B writes a letter to the editor noting that the 95 per cent confidence interval (−3.2 to 11.2) was very wide, i.e., the trial had low power relative to differences he would deem clinically significant, and that larger trials are needed to show the difference he is sure exists. The experiment, in the end, is judged inconclusive.

Evidential Approach—The investigator doing this trial makes an explicit judgment about the minimum treatment difference that he would deem clinically important, because this is necessary both for the LR calculation and his own assessment of the importance of the result. He feels that the minimum clinically significant blood pressure drop is 10 points. He only has enough money to study 30 patients in each group. The difference of four points produces an LR $(H_0 \mid H_1) = 2.1$, Support = 0.75. We have weak evidence in favor of the "no difference" hypothesis (H_0) vs the 10 point difference (H_1).

One reviewer of the study feels that a five-point drop is clinically important, and notes that the evidence favors this hypothesis (Support($\mu = 0$ vs. $\mu = 5$) ≈ -1) even though he acknowledges that the evidence is still weak. He corresponds with the author and they discuss the bases for these two different clinical judgments.

II. (N = 60) P-value Approach—Scientist A decides to redo the experiment with more people, this time in consultation with a statistician and with very high power, 97 per cent, so he is sure his trial will "detect" a difference of at least 10 points if there is one. He measures the same four-point difference that the first study did, and the p-value drops from .27 to .12. To him this represents more evidence against H_0 than in the first study, and he is annoyed that he did not enroll just a few more people to reach "significance". The statistician makes the unwelcome point that the scientist already used up his allotted alpha = .05 error in the first experiment, and this second one could not have produced an overall significant result no matter what it showed.[9] The 95% CI is −1.1 to 9.1.

Evidential Approach—The support in favor of H_0 vs H_1 increases from 0.75 to 1.5, moderate evidence for no effect vs a 10-point drop. We have more evidence *for* no difference, not against it, as the p-value seemed to indicate. (Note that with twice as many people showing exactly the same difference, the support measures exactly twice as much evidence. The support for $\mu = 0$ vs $\mu = 5$ would also double, now equaling −2.) We can add the evidence from the previous experiment for a net total of 2.25 (.75 + 1.5) units of evidence for no effect vs a 10-point drop.

III. (N = 120) P-value Approach—Firing the statistician, and doubling the sample size again, scientist A observes the same four-point difference, which is now significant at p = .028. He rejects the null hypothesis and publishes in a reputable journal. He does not mention the previous studies lest he be accused of "multiple looks". Someone writes a letter to the editor complaining that this statistically significant difference is not necessarily clinically significant. The 95% CI is 0.42 to 7.6.

Evidential Approach—The LR=20, with support = 3, strong evidence of no difference vs a 10-point difference. This investigator publishes an interpretation opposite to the one above in a competing reputable journal. The previous two studies are cited as additional supporting evidence.

Discussion

In this example we focused on the evidence for one particular alternative—a 10-point drop. In Figure 2 we see that the alternative for which there is maximum evidence is the observed four-point difference. As mentioned in the first scenario, we may want to look at the evidence for other alternatives. Or, we may want to look at an average likelihood calculated over a range of hypotheses and compare it to the likelihood at the null. In this way, the likelihood approach encourages data exploration. Which alternative to focus on

for inference depends on scientific judgment. If differences on the alternative cannot be resolved, it becomes clear why the interpretation of the data will be different. In this way LRs prevents the debate about the meaning of the "evidence' from becoming a technical statistical argument (e.g., about power), and translates it into an area that the investigators have direct experience with—the minimum important difference.

Even though the investigator may be interested only in a narrow range of alternatives, the entire result, as Poole suggested, should always be reported. This result can be represented by either the complete support curve (Figure 2), or a four to five-point summary.[19] When using a Gaussian distribution in a fixed sample size experiment, the point estimate with a 95 per cent CI is akin to a three-point summary: the point that gets highest support and the two points that receive about two units less support than the maximum (the two-unit support interval).[13] Because the experiments above satisfied those requirements, appeals to CIs produced interpretations qualitatively similar to those achieved by using likelihood. When there are multiple looks or the data are not Gaussian, the correspondence with the likelihood can be weakened or lost.

Even when two people agree on the amount of evidence, their choice of what to *believe* might still be different depending on their separate assessments of the biologic plausibility of the measured effect. Finally, what to *do*, (i.e., prescribe the drug or not) would depend on their own complex calculation of the various tradeoffs (utilities) involved; side effects, efficacy, compliance, etc. Many of these judgments are present in standard applications of p-values (in the form of asssumptions and choice of error probabilities), but they are so automated that most people are unaware they are being made.

Link between P-Values and Evidence

In the preceding example we saw that for any result, we could calculate an LR and a p-value. The power of the study set the correspondence between the two. The reason low p-values seem to mean something in practice is that in the range of 70–95 per cent power, where most experiments are done, low p-values (<.02) correspond to moderate-strong support (2–2.5 units) for the alternative. (The alternative is defined here as the "minimum difference" used in the power calculation if the observed difference is less than that. If it is greater, than the alternative can be simply the difference we observe.) The support drops below that when the power gets higher, because the same p-value represents a smaller effect as we increase the sample size.

This phenomenon is shown in Figure 2, where the sample sizes of 30, 60, and 120 correspond to powers of 78, 97, and 99.9 per cent, respectively, and the p = 0.05 point is marked in each instance. The p = .05 point never corresponds to more than moderate support for the alternative (vs the null), and it becomes very strong evidence *for the null* when the power is very high. This is why some people say that it is possible for a trial to be "too powerful", although we now see that this reflects a problem with the p-value, not the size of the trial. If we use the LR , the larger the trial, the more evidence we will measure, and the issue of being "too powerful" does not exist. P-values have a crude correspondence with evidence because of the sample sizes we tend to use, but we cannot reliably interpret a p-value as evidence until we translate it into an LR.

Conclusions

We hope this essay makes it possible to view the issues underlying the debate about statistical measures in a different light. Walker, Thompson, and Poole appealed to many likelihood concepts, but were constrained by a non-likelihood technology in making their recommendations. Walker focused mostly on the scientific context and methods with which the data is gathered,[2] which are critical to know what to *believe* on the basis of the evidence. He also downplayed the importance of adjustments for multiple comparisons, which use of the LR achieves. Thompson focused mainly on the way we should represent data—as confidence intervals.[3] By pointing us away from rigid dependence on p-values, and toward effect magnitude, he faced us more in the direction of an evidential concept. He also sensed that a p-value "substantially below .05 . . . provides potentially useful information", which the LR often bears out.

Fleiss, while concerned about some of the "abuses" p-values are prone to, is loathe to abandon the only objective standard he sees for data analysis[1]. It should be clear from this paper that p-values are neither completely objective nor the only alternative. His advocacy of "cautious"interpretation of p-values is not adequate protection from the vagaries of that index. We should be wary of any technology, however cautiously used, that fosters the illusion that we can make an inferential leap from data to conclusions without explicitly including judgment.

Poole clearly came the closest to embracing an evidential perspective in his complete rejection of CIs and single p-values, and discussion of the decision making technologies vs those that provide information for decision makers.[4] He mentioned that a complete likelihood function can be presented, but did not explain the profoundly different philosophy behind its use. He offered as a solution the complete p-value curve, with the idea that this tells us in a non-prescriptive way what parameter values are supported by the data. This is in the spirit of the likelihood approach, but it suffers from all the problems of interpreting p-values as evidence:

- It does not solve the problem of p-value dependence on the stopping rule and other comparisons (and the consequent incompatibility with Bayes Theorem).
- Using p-values limits us to thinking that our belief in a hypothesis can only be weakened by the data, not also strengthened by it.
- There is no standard way to compare parameters having different p-value "support". This makes the interpretation of a highly significant small effect in a large trial particularly problematic.
- P-values overstate the degree of conflict with the null hypothesis.

In the approaches recommended by Thompson, Walker, Poole and others for the use of p-values or CIs, they are attempting to give them some of the properties that the likelihood ratio (LR) already has. This is part of an effort to make evidential sense out of indices not designed to measure evidence. It is hoped that a better understanding of likelihood methods and the philosophy of inductive inference will make thought and discussion about the meaning of observed patterns an integral part of the statistical process, and move us permanently away from the notion that in the data lie absolute proofs and truths that statistical technologies can reveal. Unlike the p-value, the use of evidential measures forces us to bring scientific judgment to data analysis, and

shows us the difference between what the data are telling us and what we are telling ourselves.

ACKNOWLEDGMENTS

Dr. Goodman is supported by NCI NRSA grant #5-T32-CA 09314. Parts of this paper were presented at the meeting of the Society for Clinical Trials, Atlanta, GA, in May 1987.

APPENDIX

Derivation of Figure 2

The systolic blood pressures of the compared populations were set as having a $\sigma^2 = 200$, thus the distribution of the *difference* has a $\sigma_p^2 = 2*200/N$, or $\sigma_p = 20/\sqrt{N}$. The support of a drop of 10, compared to no effect, at an observed difference of four points, is:

$$[1]\ \log(e^{-(\bar{x})^2/(2*400/N)}/e^{-(\bar{x}\ -\ 10)^2/(2*400/N)}) = (10^2 - 2*10*4)*N/(2*400) = N/40.$$

The Z-score is:

$$[2]\ Z = 4/\sigma_p = 4\sqrt{N/20} = \sqrt{N/5}.$$

The two-sided p-value is then calculated from the Z-score.

To calculate the support corresponding to a given p-value, the Z-score corresponding to that p-value is obtained, the actual difference this Z score represents ($\bar{x} = Z*\sigma_p$) is calculated, and this $\bar{x}$ is substituted into Equation 1.

REFERENCES

1. Fleiss JL: Significance tests have a role in epidemiologic research: reactions to A.M. Walker. (Different Views) Am J Public Health 1986; 76:587.
2. Walker AM: Reporting the results of epidemiologic studies. (Different Views) Am J Public Health 1986; 76:556–558.
3. Thompson WD: Statistical Criteria in the interpretation of epidemiologic data. (Different Views) Am J Public Health 1987; 77:191–194.
4. Poole C: Beyond the confidence interval. (Different Views) Am J Public Health 1987; 77:195–199.
5. Hacking I: The Logic of Statistical Inference. Cambridge: Cambridge University Press, 1965.
6. Popper KR: The Logic of Scientific Discovery. New York: Harper and Row, 1959.
7. Hempel CG: Studies in the Logic of Confirmation, in Aspects of Scientific Explanation. New York: Free Press, 1965.
8. Carnap R: The Concept of Confirming Evidence, in The Logical Foundations of Probability, 2nd Ed. Chicago: University of Chicago Press, 1962.
9. Cornfield J: Sequential trials, sequential analysis, and the Likelihood Principle. Am Statist 1966; 20:18–23.
10. Jeffreys H: Theory of Probability (3rd Ed). Oxford: Oxford University Press, 1961.
11. Birnbaum A: On the Foundations of Statistical Inference. JASA 1962; 298:269–306.
12. Fisher RA: Statistical Methods and Scientific Inference. London: Oliver and Boyd, 1956.
13. Edwards AWF: Likelihood. Cambridge: Cambridge University Press, 1972.
14. Hodges A: Alan Turing: The Enigma. New York: Simon and Schuster, 1983; 197.
15. Berger JO: Are P-values reasonable measures of accuracy? In: Francis IS, Manly BFJ, Lam FC (eds): Pacific Statistical Congress, Elsevier, (North-Holland), 1986.
16. Berger JO, Sellke T: Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence. JASA 1987; 82:112–139.
17. Shafer G: A Mathematical Theory of Evidence. Princeton: Princeton University Press, 1976.
18. Good IJ: Probability and the Weighing of Evidence. New York: Charles Griffin & Co, 1950.
19. Barnard GA: The Use of the Likelihood Function in Statistical Practice. Proc V Berkeley Symposium 1966; 1:27–40.

Kellogg Foundation Funds 3-Year 'Healthy Cities Indiana' Project

The W. K. Kellogg Foundation recently awarded nearly a half-million-dollar grant to the Indiana University and the Indiana Public Health Association to fund *Healthy Cities Indiana,* a three-year project to support six cities in their efforts to promote a healthier life for their citizens and develop solutions to community health problems. The six cities will be selected from those around the state which express a serious interest in the project.

The five major elements of *Healthy Cities Indiana* are:

- Assessment of current city health status,
- Formulation and adoption of action-based city health plans,
- Development of solutions to problems on a community-wide basis,
- Mutual support collaboration and shared learning among the six cities,
- Sharing of information with policy makers and others interested in healthy cities.

Dr. Beverly Flynn, professor and chair of the University's Department of Community Health Nursing, will serve as a project director of the $464,200 grant which will fund workshops for city officials, provide technical support from the Department of Community Health Nursing faculty and staff, and help the cities develop a network for information exchange and shared learning experiences.

According to Dr. Flynn, the complex problems facing urban areas need to be worked out through intense cooperation between health professionals and community leaders. "Factors such as unemployment, poor housing, access to health care, AIDS, and environmental and occupational conditions continue to affect the health of our citizens," she said. It is through the process "of working together that appropriate solutions to problems can be found," Dr. Flynn emphasized.

The Healthy Cities movement is widespread in Canada and Europe," she said, but has not been demonstrated in the US. Dr. Flynn has met with project leaders in Canada and England and feels that much can be learned from those experiences. In the *Healthy Cities Indiana* project, community leaders will become involved in learning new concepts, understanding broad categories of data, and blending these concepts and data into new solutions to community health problems. The project will continue through July 1991. For further information about the project, contact Dr. Flynn or Melinda Rider, associate project director, at Indiana University School of Nursing, 610 Barnhill Drive, Indianapolis, IN 46223. Tel: (317) 274-2129.