Neyman, J. 1976. Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena. *Comm. Stat. Theor. Methods* 8:737–751.

Pearson, E. S. 1947. The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table. *Biometrika* 34:139–167. Reprinted in Pearson (1966).

Pearson, E. S. 1950. On Questions Raised by the Combination of Tests Based on Discontinuous Distributions. *Biometrika* 37:383–398. Reprinted in Pearson (1966).

Pearson, E. S. 1955. Statistical Concepts in Their Relation to Reality. *J. Roy. Stat. Soc.*, ser. B, 17:204–207.

Pearson, E. S. 1966. *The Selected Papers of E. S. Pearson.* Berkeley: University of California Press.

Pearson, E. S., and Neyman, J. 1930. On the Problem of Two Samples. *Bull. Acad. Pol. Sci.*, 73–96. Reprinted in Neyman and Pearson (1967).

Royall, R. M. 1997. *Statistical Evidence: A Likelihood Paradigm.* London: Chapman and Hall.

Royall, R. M. 2004. The Likelihood Paradigm for Statistical Evidence. Chapter 5 in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Empirical, Statistical, and Philosophical Considerations.* Chicago: University of Chicago Press.

Savage, L. J. 1972. *The Foundations of Statistics.* New York: Dover.

Spanos, A. 2000. Revisiting Data Dining: "Hunting" with or without a License. *J. Econ. Methodology* 7:231–264.

5 The Likelihood Paradigm for Statistical Evidence

Richard Royall

ABSTRACT

Statistical methods aim to answer a variety of questions about observations. A simple example occurs when a fairly reliable test for a condition or substance, C, has given a positive result. Three important types of questions are: (1) Should this observation lead me to believe that C is present? (2) Does this observation justify my acting as if C were present? (3) Is this observation evidence that C is present? We distinguish among these three questions in terms of the variables and principles that determine their answers. Then we use this framework to understand the scope and limitations of current methods for interpreting statistical data as evidence. Questions of the third type, concerning the evidential interpretation of statistical data, are central to many applications of statistics in science. We see that for answering them current statistical methods are seriously flawed. We find the source of the problems, and propose a solution based on the law of likelihood. This law suggests how the dominant statistical paradigm can be altered so as to generate appropriate methods for (i) objective representation and measurement of the evidence embodied in a specific set of observations, as well as (ii) measurement and control of the probabilities that a study will produce weak or misleading evidence.

INTRODUCTION

An important role of statistical analysis in science is interpreting observed data as evidence—showing "what the data say." Although the standard statistical methods (hypothesis testing, estimation, confidence intervals) are routinely used for this purpose, the theory behind those methods contains

no defined concept of evidence and no answer to the basic question "When is it correct to say that a given body of data represents evidence supporting one statistical hypothesis over another?" or to its sequel, "Can we give an objective measure of the *strength* of statistical evidence?" Because of this theoretical inadequacy, the use of statistical methods in science is guided largely by convention and intuition and is marked by unresolvable controversies (such as those over the proper use and interpretation of *P*-values and adjustments for multiple testing).

We argue that the law of likelihood represents the missing concept and that its adoption in statistical theory can lead to a frequentist methodology that avoids the logical inconsistencies pervading current methods while maintaining the essential properties that have made those methods into important scientific tools.

STATISTICAL EVIDENCE

By "statistical evidence," we mean observations that are interpreted under a probability model. The model consists of a collection of probability distributions, and the observations are conceptualized as having been generated from one of the distributions.

For example, a subject is given a diagnostic test for a disease and the result is positive. This observation might be interpreted as a realization of a random variable X whose possible values are 1 (positive) and 0 (negative). The distribution of X is determined by whether the subject does or does not have the disease, as shown in the following table of probabilities.

	Test Result	
	$X = 1$	$X = 0$
disease present	.94	.06
disease absent	.02	.98

This simple model has only two probability distributions (given in the two rows of the table). In this context, the observed test result is an example of statistical evidence.

Three Questions

Statistics is the discipline concerned with statistical evidence—producing, modeling, interpreting, communicating, and using it. We will focus on an

area of statistics that is central to its role in science, interpreting and communicating statistical evidence per se. To distinguish this problem area from some other branches of statistics, and to introduce its essential principle, we consider the above diagnostic test, and three conclusions about disease status that might be appropriate after a positive result has been observed:

This person probably has the disease.
This person should be treated for the disease.
This test result is evidence that this person has the disease.

How can we determine which, if any, of these conclusions are correct?

The first is a statement about the present state of uncertainty concerning the subject's disease status, i.e., the conditional probability of disease, given the positive test result. It states that this probability is greater than .5. The above model does not determine whether this conclusion is true. This is because the conditional probability of disease (given the positive test result) depends not only on the probabilities that comprise the model, but also on a quantity that is not represented in this model. That missing quantity is the probability of disease before the test (the prior probability). Precisely *how* the present uncertainty depends on the prior is detailed in elementary probability theory by Bayes' theorem. For example, if the test was used in a mass screening program for a rare disease, and if the subject is simply one of those whose results were positive, then the prior probability is the prevalence of the disease in the screened population. And if that probability is less than .021, then Bayes' theorem shows that the present probability of disease is still less than .5, so that, although the test is positive, the first conclusion is wrong—this person probably does *not* have the disease. But if, instead of an anonymous participant in the screening program, she is a patient whose symptoms implied a disease probability of .10 before the test, then the probability of disease is now .84, and the first conclusion is quite correct.

The correctness of the second statement ("This person should be treated for the disease") depends on the present probability of disease, so that it too depends on the prior probability. But it depends on other factors as well, such as the costs of treating and of not treating, both when the disease is present and when it is not. Even when the first conclusion is wrong, the second might be correct. This would be true if the treatment is a highly effective one, with little cost or risk, while not treating a patient with the disease is disastrous. But under different conditions of prior uncertainty and costs, the opposite conclusions might be appropriate—it might be best that the subject *not* be treated, even though she probably has the disease.

The third conclusion, unlike the first two, requires for its appraisal nothing more than the probability model represented by the table. Under that model, the third conclusion is correct, regardless of the disease probability before the test and regardless of the costs associated with whatever treatment decisions might be made—the positive test result *is* evidence that this person has the disease. This strongly intuitive conclusion is certified by the basic rule for interpreting statistical evidence, which will be stated in the next section.

Each of the three conclusions represents an answer to a different question:

What should I believe?
What should I do?
How should I interpret this body of observations as evidence?

These three questions define three distinct problem areas of statistics.

It is the third problem area, proper interpretation of statistical data as evidence, that we are concerned with in this paper. It is a critical question in scientific research, and, as the simple diagnostic test example shows, it is the only one of the three questions that can be answered independently of prior beliefs.

The Law of Likelihood

We have seen that, after the positive test result has been observed, the conclusion that the subject probably does not have the disease is appropriate when the prior probability is small enough. Similarly, the conclusion that the best course of action is not to treat for the disease is appropriate under certain conditions on the prior probability and the potential costs. But to interpret the positive test result as *evidence that the subject does not have the disease* is never appropriate—it is simply and unequivocally wrong. Why is it wrong?

The interpretation is wrong because it violates the fundamental principle of statistical reasoning. That principle, the basic rule for interpreting statistical evidence, is what Hacking (1965, 70) named the law of likelihood. It states:

> If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence.

This says simply that if an event is more probable under hypothesis A than hypothesis B, then the occurrence of that event is evidence supporting A over B—the hypothesis that did the better job of predicting the event is better supported by its occurrence. It further states that the *degree* to which occurrence of the event supports A over B (the strength of the evidence) is quantified by the ratio of the two probabilities.

When uncertainty about the hypotheses, before $X = x$ is observed, is measured by prior probabilities, $P(A)$ and $P(B)$, the law of likelihood can be derived from elementary probability theory. In that case, the quantity $p_A(x)$ is the conditional probability that $X = x$, given that A is true, $P(X = x|A)$, and $p_B(x)$is $P(X = x|B)$. The definition of conditional probability implies that

$$\frac{P(A|X = x)}{P(B|X = x)} = \frac{p_A(x)P(A)}{p_B(x)P(B)}.$$

This formula shows that the effect of the statistical evidence (the observation $X = x$) is to change the probability ratio from $P(A)/P(B)$ to $P(A|X = x)/P(B|X = x)$. The likelihood ratio, $p_A(x)/p_B(x)$, is the exact factor by which the probability ratio is changed. If the likelihood ratio equals 5, then the observation $X = x$ constitutes evidence just strong enough to cause a fivefold increase in the probability ratio. Note that the strength of the evidence is independent of the prior probabilities. (The same argument and conclusion apply when $p_A(x)$ and $p_B(x)$ are not probabilities but probability densities at x.)

The likelihood ratio is a precise and objective numerical measure of the strength of statistical evidence. Practical use of this measure requires that we learn to relate it to intuitive verbal descriptions such as "weak," "fairly strong," "very strong," etc. The values 8 and 32 have been suggested as benchmarks for likelihood ratios—observations with a likelihood ratio of 8 (or 1/8) constitute moderately strong evidence, and observations with a likelihood ratio of 32 (or 1/32) are strong evidence. These benchmark values come from considering the various possible results of one of the simplest of experiments (Royall, 1997) and are similar to others that have been suggested (Jeffreys, 1961; Edwards, 1972; Kass and Raftery, 1995).

Misleading Evidence

The positive result on our diagnostic test, with a likelihood ratio (LR) of .94/.02 = 47, constitutes strong evidence that the subject has the disease. This interpretation of the test result is correct, regardless of that subject's actual disease status. If she does not have the disease, then the evidence is mis-

leading. We have not made an error—we have interpreted the evidence correctly. *It is the evidence itself that is misleading.*

Statistical evidence, properly interpreted, can be misleading. But we cannot observe strong misleading evidence very often. In our example, if the disease is not present table 1 shows that the probability of a (misleading) positive test is only .02. It is easy to prove that in other situations the probability of observing misleading evidence this strong or stronger (LR $\geq$ 47) can be slightly greater, but it can never exceed 1/47 = .0213. We can state a *universal bound on the probability of misleading evidence:* If hypothesis A implies that the probability that a random variable X has one probability density (or mass) function, $f_A(\cdot)$, while hypothesis B implies another, $f_B(\cdot)$, then if A is true the probability of observing evidence supporting B over A by a factor of k or more cannot exceed $1/k$. That is, $P_A(f_B(X)/f_A(X) \geq k) \leq 1/k$ (Royall, 1997).

This bound has been noted by various authors (e.g., Smith, 1953; Birnbaum, 1962). Much tighter bounds apply in important special cases. For example, when the two distributions are normal with different means and a common variance, the universal bound $1/k$ can be replaced by the much smaller value, $\Phi(-\sqrt{2\ln(k)})$, where Φ is the standard normal distribution function. In that case the probabilities of misleading evidence, for the proposed benchmarks $k = 8$ and $k = 32$ (representing "pretty strong" and "strong" evidence respectively), cannot exceed .021 and .004 (Royall, 1997).

The universal bound $1/k$ applies even if we deliberately seek evidence supporting B over A by continuing to make observations on X until our sample gives a likelihood ratio of at least k in favor of B. When A is true, the probability is at most $1/k$ that we will succeed, sooner or later finding that the accumulated observations represent strong evidence in favor of B. That is, the probability is at least $1 - 1/k$ that we will *never* succeed, sampling forever without once finding that our data represent strong evidence in favor of B (Robbins, 1970).

The Likelihood Function

As a second example of statistical evidence, consider observing a sequence of tosses of a 40¢ coin. The coin is asymmetric, consisting of an ordinary quarter, nickel, and dime that have been glued together so that the heads of the dime is on one side and the tails of the quarter is on the other. If we model the tosses as independent trials with a common probability of heads, θ, then every value of θ between 0 and 1 determines a different probability distribution. Under this model a sequence of observations, such as 1, 1, 0, 1, 1, 0, . . . (heads = 1, tails = 0) represents statistical evidence. The probability that a sequence of n tosses will produce observations $x_1, x_2, \ldots, x_n$, is $\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^k(1-\theta)^{n-k}$ where $k = \sum_1^n x_i$, the number of heads. For a given set of observations this probability is a function of the variable θ and is called the *likelihood function:*

$$L(\theta) = \theta^k(1-\theta)^{n-k} \qquad 0 \leq \theta \leq 1.$$

The law of likelihood gives this function its meaning: $L(\theta)$ determines the relative support for every pair of values of θ, the probability of heads. Observation of a sequence of n tosses that contains k heads represents evidence supporting the hypothesis that $\theta = \theta_1$ over the hypothesis that $\theta = \theta_2$ by the factor $L(\theta_1)/L(\theta_2)$.

Two likelihood functions that differ only by a constant multiple are equivalent, because they give identical likelihood ratios for all pairs of values of the parameter. That is, the likelihood function is defined only up to an arbitrary constant multiplier.

To produce some evidence about the probability of heads for my 40¢ piece, I performed a simple experiment. I tossed the coin 17 times and noted the results. Heads appeared 9 times. To see what these observations say about the probability of heads, we look at the likelihood function, $\theta^9(1-\theta)^8$, which is shown in figure 5.1. It is a graphical representation of this statistical evidence. We have scaled this function so that its maximum value is 1. We have noted the value of θ that is best supported by the observations (9/17 = .53), as well as the 1/8 and 1/32 likelihood intervals (LIs).

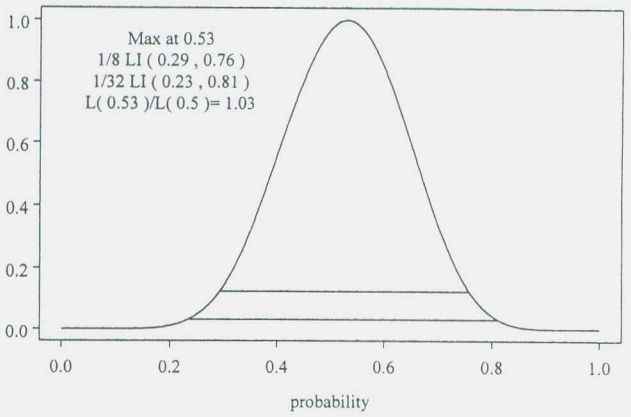

FIGURE 5.1 Likelihood for probability of heads with 9 heads observed in 17 tosses.

These intervals, where the scaled likelihood is at least 1/8 (1/32), represent the values of θ that are consistent with the observations at the levels of the two benchmarks, 8 and 32—if a value is in the 1/8 interval, (0.29, 0.76), then there is no alternative that is better supported by a factor of 8 or more (fairly strong evidence).

The Likelihood Principle

Suppose two instances of statistical evidence generate the same likelihood function. According to the law of likelihood, this means that for every pair of parameter values, θ_1 and θ_2, the strength of the evidence in support of θ_1 vis-à-vis θ_2, $L(\theta_1)/L(\theta_2)$, is the same in both instances. Both will have the same impact on any prior probability distribution for θ. In this powerful sense, they are equivalent. On the other hand, if two instances of statistical evidence are equivalent, then all of the likelihood ratios, $L(\theta_1)/L(\theta_2)$, must be equal, which means that the two likelihood functions are the same. Therefore, the law of likelihood implies that two instances of statistical evidence are equivalent if and only if they generate the same likelihood function. This proposition is called the likelihood principle (Birnbaum, 1962; Edwards, Lindman, and Savage, 1963). It means that *the likelihood function is a mathematical representation of the statistical evidence per se.*

Our experiment with the 40¢ piece produced some observations. The likelihood function shows what those observations mean, as evidence about the probability of heads, in the context of the model that we are using. It also shows that, within this context, certain aspects of the observations have no effect on their evidential meaning and are, in that sense, irrelevant. For example, I actually observed a particular sequence of heads and tails: 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1. Other observations, consisting of 9 heads and 8 tails in some different order, would give the same likelihood function, $\theta^9(1-\theta)^8$. Those observations would constitute evidence of exactly the same strength as ours in support of any value, θ_1, versus any other, θ_2, and would therefore represent evidence that is equivalent to ours. The meaning of the observations, as evidence about the probability of heads, depends only on the number of heads and the number of tails. The order of the observations does not affect any of the likelihood ratios (i.e., it does not affect the likelihood function)—the order is irrelevant.

Besides showing that certain characteristics of the observations are irrelevant to their proper interpretation as evidence, the likelihood principle reveals that certain aspects of the experimental procedure that produced those observations are also irrelevant. My observation of 9 heads in 17 trials might have been generated by fixing the number of trials at 17, observing the (random) number of heads, say H, and finding $H = 9$. But it might have been produced in another way; instead of stopping after a fixed number of trials, I could have generated these observations by fixing the number of heads to be observed at 9, observing the (random) number of trials, say N, required to produce 9 heads, and finding $N = 17$. The probability of observing 9 heads in seventeen trials is $P(H = 9) = \binom{17}{9}\theta^9(1-\theta)^8$ in the first experiment, and $P(N = 17) = \binom{16}{8}\theta^9(1-\theta)^8$ in the second. Although these probabilities differ, the likelihood functions are the same (proportional to $\theta^9(1-\theta)^8$). Thus, it does not matter which procedure I actually used (stopping after 17 trials, or after 9 heads). What these observations mean, as evidence about the probability of heads, is shown in figure 5.1: it is the same in both cases. The stopping rule does not affect the likelihood function—for interpreting the observed data as evidence, the stopping rule is irrelevant.

This conclusion, irrelevance of the stopping rule, implies that conventional frequentist statistical methods are not appropriate *for interpreting the observed data as evidence.* This is because those methods, when applied to our observations (9 heads in 17 trials) give different results under the two stopping rules. For example, the observed proportion of successes, 9/17, is an unbiased estimator of θ under the first stopping rule, but not under the second. Furthermore, this estimator has different standard errors under the two stopping rules. Confidence coefficients and P-values also differ, depending on which stopping rule was used. P-values do not measure the strength of the evidence, because the evidence is the same in both cases, but the P-values (for testing the hypothesis that θ is .8 vs. some smaller value, for example) are different. Confidence intervals do not show "what the data say" about θ, because the data say the same thing in both cases, while the confidence intervals are different.

We will consider this example and the radical conclusion that it illustrates in more detail in the next section. The irrelevance of the stopping rule, in the evidential interpretation of observed data, will be illustrated again below.

STATISTICS IN SCIENCE: THE MISSING LINK

Statistics today is in a conceptual and theoretical mess. The discipline is divided into two rival camps, the frequentists and the Bayesians, and neither camp offers the tools that science needs for objectively representing and in-

terpreting statistical data as evidence. In this section, we will identify the theoretical deficiency that has produced this methodological one, and see how to correct both.

Bayesian statistics is primarily concerned with the question of how one's beliefs should change in response to new statistical evidence; that is, its focus is on the first of the three questions listed above ("What should I believe?"). As a leading Bayesian put it, "The main subject matter of statistics is the study of how data sets change degrees of belief; from prior, by observation of A, to posterior. They change by Bayes' theorem" (Lindley, 1965, 30). The result of a Bayesian statistical analysis is a (posterior) probability distribution for the parameter. Now, the posterior distribution is determined by the prior and the likelihood function *together* (as explained by Bayes' theorem). Therefore, as shown in the diagnostic test example above, a Bayesian analysis requires, in addition to the probability model for the observed data (which determines the likelihood function), a (prior) probability distribution for the parameter. A frequentist model for my observations on the 40¢ coin represents the tosses as independent trials with some unknown probability, θ, of heads. A Bayesian model must supplement this with a prior probability distribution for θ. The two camps use the same model for the probability distribution of the observable random variables (the results of the tosses); but the Bayesian requires, in addition, a prior probability distribution for the parameter. This prior distribution represents the experimenter's state of uncertainty about the parameter before the observations are made (Edwards, Lindman, and Savage, 1963).

The observations affect the posterior probability distribution only through the likelihood function: for a given prior distribution, if different observations (perhaps from different experiments) produce the same likelihood function, then they produce the same posterior probability distribution. Thus, the Bayesian approach leads inevitably to the likelihood principle. But while they have embraced the principle, Bayesians have shown little interest in likelihoods per se, i.e., likelihoods without prior probabilities. Savage himself wrote (1962, 307), "I, myself, came to take personalistic statistics, or Bayesian statistics as it is sometimes called, seriously only through recognition of the likelihood principle. I suspect that once the likelihood principle is widely recognized, people will not long stop at that halfway house." More recently, Berger and Wolpert (1988, 124) argued that "sensible use of the likelihood function seems possible only through Bayesian analysis," and Lindley (1992, 415) went so far as to proclaim that "the only satisfactory measures of support are probability-based. Likelihood will not do."

Bayesian efforts to show what the data say have concentrated on (i) searching for prior probability distributions to represent total ignorance (the state of knowledge of an ideal ignoramus) or, failing that, (ii) arguing that certain distributions should be adopted as standard "reference priors" (Bernardo, 1979; Berger and Bernardo, 1992). Posterior probability distributions corresponding to such priors would then supposedly show what the data say. That is, Bayesians have tried to use posterior probability distributions, appropriate to question 1 ("What should I believe?") to answer question 3 ("How should I interpret this body of observations as evidence?"). Non-Bayesians generally judge these attempts to have failed (e.g., Edwards, 1969; 1992, sec. 4.5). The reason for this failure is found in the law of likelihood—they have failed because probabilities represent and measure degrees of belief or uncertainty, not evidence. The law of likelihood reveals that *evidence has a different mathematical form than uncertainty.* It is likelihood ratios, not probabilities, that represent and measure statistical evidence (Royall, 1997, secs. 1.13, 8.6). It is the likelihood function, and not any probability distribution, that shows what the data say.

Science, looking to statistics for objective ways to represent and quantify evidence, has not embraced Bayesian methods. One obvious reason is the need for prior probabilities in Bayesian analyses. Since these probabilities are usually personal, or subjective (the quest for objective "ignorance" priors or widely acceptable "reference" priors having failed), they are widely seen as incompatible with the scientific need for objectivity. As Efron (1986, 4) put it, "The high ground of scientific objectivity has been seized by the frequentists."

But all is not well with statistics in science. As we saw in the previous section, there is something fundamentally wrong with today's standard frequentist methodology for evidential interpretation of scientific data. Allan Birnbaum (1970), while advocating the use of this methodology, acknowledged that it consists of "an incompletely formalized synthesis of ingredients borrowed from mutually incompatible theoretical sources."

The theoretical problems have created practical ones, as evidenced by the endless controversy over the proper use and interpretation of statistical hypothesis testing in science (Morrison and Henkel, 1970; Cohen, 1994; Bower, 1997; Thompson, 1998; Goodman, 1998; Sterne and Smith, 2001). This controversy flows from the discord between theory and practice—frequentist statistical *theory* is based on Neyman's (1950) behavioristic view that a hypothesis test is a procedure for choosing between the two hypotheses (with evidential interpretation explicitly disallowed), while science's main use of hypothesis tests is for showing the direction and strength of statisti-

cal evidence. The theory is aimed at our question 2 ("What should I do?"), but the methods are used to answer question 3. (For further discussion and details see Royall, 1997, chaps. 2, 3.)

The frequentists' claim to occupy the high ground of objectivity has even been challenged. In fact, as Edwards, Lindman, and Savage (1984, 59) pointed out, dependence on stopping rules makes conventional frequentist statistical procedures (which they referred to as "classical procedures") subjective: "The irrelevance of stopping rules is one respect in which Bayesian procedures are more objective than classical ones. Classical procedures . . . insist that the intentions of the experimenter are crucial to the interpretation of data." These authors illustrated their point with an example like our observation of 9 heads in 17 tosses of a 40¢ piece. Suppose you and I were collaborators on that experiment, but we disagreed about which stopping rule should be used. I wanted to fix the number of tosses (at 17), but you wanted to fix the number of heads (at 9). Instead of postponing the experiment until we could agree on a stopping rule, we decided to begin sampling and to continue so long as both stopping rules said that we should (i.e., until we had either 9 heads or 17 observations, whichever came first). At that point we would decide which rule to apply. Because the ninth head occurred on the seventeenth toss (so *both* rules told us to stop), the decision was never made. But our data cannot be analyzed using conventional frequentist methods until it is made, i.e., until someone determines *whose will would have prevailed if different results had been observed*—if the ninth head had occurred earlier, would you have persuaded me to stop the study, or would I have persuaded you to continue? If, after 17 tosses, we had not yet observed 9 heads, would I have persuaded you to stop? Although both stopping rules assign exactly the same probability, $\theta^9(1-\theta)^8$, to the actual observations, (1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1), conventional frequentist methods imply, quite wrongly, that what these observations *mean*, as evidence about the tendency of the coin to fall heads, depends on whose will would have prevailed, yours or mine, if different observations had been made.

Why does science continue to use frequentist methods, despite their well-known logical defects? We have seen one important reason—the available alternative, Bayesian statistics, has a more conspicuous subjective component. But there is another, more practical reason: Science has embraced these methods because they provide something that science needs. Specifically, *frequentist statistical methods provide explicit objective measure and control of the frequency of errors (and Bayesian methods do not)*. This is illustrated most clearly in the formulas that are routinely used to determine sample size when a scientific study is planned. The study is conceptualized as a procedure for choosing between two hypotheses, with potential errors of two types. Probabilities of the two types of error are set at specified target levels. These are plugged into a formula that gives the required sample size. The researcher cannot eliminate the *possibility* of errors. But by using frequentist statistical methods, he can calculate and control their *probability*.

Can statistics avoid the logical inconsistencies that pervade current frequentist methods for interpreting data as evidence, while still providing what science requires—objective measure and control of the risk of unsatisfactory results? Yes. The key to accomplishing these two goals is to recognize that probabilities, which properly measure the uncertainties associated with a given procedure for generating statistical evidence, are not appropriate for measuring the strength of evidence produced. As Fisher (1959, 93) put it, "As a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence."

We must distinguish between the strength of evidence and the probability that a procedure will produce evidence of a given strength. The problem with current frequentist theory is that, lacking an explicit concept of evidence, it attempts to use the same quantities (probabilities) to measure both the chance of errors and the strength of observed evidence.[1] The solution, presented in the next section, is found in the law of likelihood, which embodies the explicit, objective, quantitative concept of evidence that is missing from current frequentist theory, and which explains that it is likelihood ratios, not probabilities, that measure evidence. Probabilities measure uncertainty; likelihood ratios measure evidence.

TWO APPROACHES TO PLANNING A SIMPLE EXPERIMENT

In this section, we consider the problem of deciding how many observations will be made in a scientific study or experiment. The Neyman-Pearson formulation and solution to this problem represents the paradigm that guides modern frequentist statistical theory (Royall, 1997, chap. 2). We first describe that paradigm in its simplest, clearest form. Then we see how it is changed when the problem is reformulated in terms of statistical evidence.

1. For further discussion, see Royall (1997, chap. 5).

Neyman-Pearson Paradigm

Purpose: The experiment is a procedure for choosing between two hypotheses. We will make some observations, and they will determine which hypothesis is chosen.

Probability model: Independent random variables $X_1, X_2, \ldots, X_n$, identically distributed as X, will be observed. The model consists of two probability distributions for X, corresponding to simple hypotheses H_1 and H_2.

Objective of statistical analysis: We wish to use the observations to choose between H_1 and H_2.

Desiderata: Noting that we can make errors of two types, choosing H_1 when H_2 is true, and vice-versa, we want:

(a) To measure and control the error probabilities. (We want to be pretty sure (probability $\geq 1 - \alpha$) that we won't choose H_2 when H_1 is true and pretty sure (probability $\geq 1 - \beta$) that we won't choose H_1 when H_2 is true.)

(b) To minimize sample size subject to (a).

The immediate goal of many scientific studies is not to choose among hypotheses, but to generate empirical evidence about them. This evidence will be communicated in reports and journal articles. Then various parties will use it, in combination with other information, as well as judgments about the consequences of alternative actions, in making a variety of choices and decisions, many that were not even imagined by the authors of the study. An epidemiological study, for example, might produce strong statistical evidence that cigarette smokers have a much greater risk of dying from a certain type of cancer than nonsmokers. The published evidence will be used by many parties—legislators, lawyers, tobacco farmers, smokers, nonsmokers, researchers in other fields (oncology, genetics, etc.), insurance companies—in making innumerable choices and decisions. From this perspective, a more realistic formulation of the problem of planning a scientific study uses the evidence-generating paradigm.

Evidence-Generating Paradigm

Purpose: The experiment is a procedure for generating empirical evidence about the two hypotheses. We will make some observations and interpret them as evidence.

Probability model: Independent random variables $X_1, X_2, \ldots, X_n$, identically distributed as X, will be observed. The model consists of two probability distributions for X, corresponding to simple hypotheses H_1 and H_2.

Objective of statistical analysis: We wish to interpret the observations as evidence regarding H_1 vis-à-vis H_2.

Desiderata: We will make no errors—we will interpret the evidence correctly. But the evidence itself can be unsatisfactory in two ways: It can be weak (evidence that does not strongly support either hypothesis), or it can be misleading (strong evidence for H_1 when H_2 is true or vice-versa). We want:

(a) To measure and control the probabilities of observing weak or misleading evidence. (If either hypothesis is true we want to be pretty sure (probability $\geq 1 - W$) that we won't find weak evidence, and pretty sure (probability $\geq 1 - M$) that we won't find strong evidence in favor of the other one, or misleading evidence.)

(b) To minimize sample size subject to (a).

To compare this evidence-generating paradigm to the preceding Neyman-Pearson decision-making paradigm, we consider the familiar example of observing normally distributed random variables with known standard deviation. Hypothesis H_1 specifies one value for the mean, and H_2 specifies another value, larger by a fixed multiple, Δ, of the standard deviation (i.e., $\mu_2 = \mu_1 + \Delta\sigma$).

First, we apply the Neyman-Pearson approach. For a given probability of the first type of error, α, the probability of the second type, β, is determined by the sample size n: $\beta(n) = \Phi(z_{1-\alpha} - n^{1/2}\Delta)$ where $z_{1-\alpha}$ is the $100(1 - \alpha)$th percentile of the standard normal distribution. The standard formula (e.g., Pagano and Gauvreau, 1993, 225) shows that the minimum sample size that will control the error rates at the target values (α, β) is $n = (z_{1-\alpha} + z_{1-\beta})^2/\Delta^2$. For example, if we take $\alpha = \beta = .05$, then for $\Delta = 1$, $n = (1.645 + 1.645)^2 = 10.8$; 11 observations are sufficient.

If we take the evidential approach, with the suggested benchmark value $k = 8$, representing pretty strong evidence, then the probability of strong misleading evidence (likelihood ratio greater than 8 in favor of the false hypothesis) is the same under both hypotheses. For a sample of n observations, this probability is: $M(n) = \Phi(-\sqrt{n}\Delta/2 - \ln(8)/(\sqrt{n}\Delta))$.

The probability of weak evidence (likelihood ratio between 1/8 and 8) is also the same under both hypotheses:

$$W(n) = \Phi(\sqrt{n}\Delta/2 + \ln(8)/(\sqrt{n}\Delta)) - \Phi(\sqrt{n}\Delta/2 - \ln(8)/(\sqrt{n}\Delta)).$$

These two functions, $M(n)$ and $W(n)$, are shown in figure 5.2 for $\Delta = 1$. Dashed lines indicate the Neyman-Pearson error probabilities, $\alpha = .05$ (constant) and β_n, for comparison. Note that $\beta_n = .05$ when $n = 10.8$, as found in the previous paragraph.

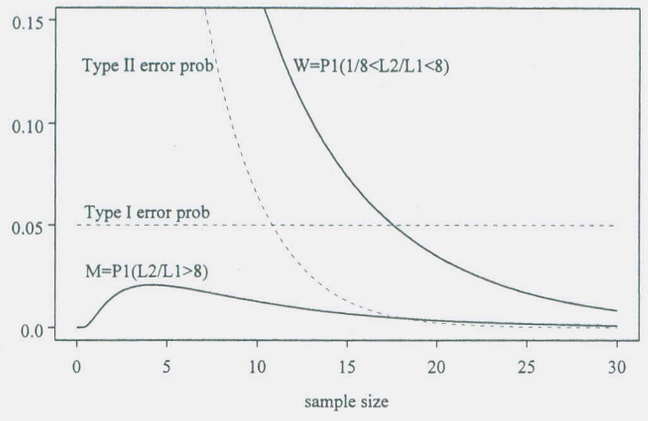

FIGURE 5.2 Probabilities of weak and misleading evidence with normal means (Differences = one standard deviation).

Several things are clear from figure 5.2.

(a) Both probabilities, M and W, can be held below any desired positive bound, no matter how small, by making enough observations.

(b) The probability of misleading evidence, M, is very small for small n, *increasing* with n until reaching a maximum (at $n = 2\ln(8)/\Delta^2$), and decreasing thereafter.

(c) The *maximum* probability of misleading evidence, over all values of n, is small. (It is actually $\Phi(-\sqrt{2\ln(8)})$, or .021.) It is attained when the sample size is so small that the probability of weak evidence is large ($W(n) = \Phi(\sqrt{2\ln(8)}) = .979$).

(d) Because of (c), the sample size calculation is driven by the constraint on the probability of weak evidence. We need large samples in order to have a good chance of getting strong evidence. *The nature of statistical evidence is such that the chances of getting misleading evidence are small at all sample sizes.* This remains true in general: for any $k > 1$, the maximum probability of misleading evidence (over all n) is $\Phi(-\sqrt{2\ln(k)})$. Even for the unreasonably small benchmark value of $k = 4$, this probability cannot exceed .05.

(e) There are essential differences between the "analogous" quantities α and M. We can fix α, the probability that when H_1 is true we will choose H_2, at any level we like. But as noted in (d) there are natural limits on M, the probability that when H_1 is true we will find strong evidence in favor of H_2.

(f) The standard calculation gives a sample size that is too small to ensure

that the experiment will, with high probability, produce strong evidence about these two hypotheses. At $n = 11$, where β, the probability of an error of the second type, falls just below .05, the probability of finding only weak evidence is about three times as great: $W(11) = .14$.

PLANNING A STUDY VS. INTERPRETING OBSERVATIONS AS EVIDENCE

The probabilities of weak and misleading evidence are important in planning a study, but they play no role in the proper interpretation of the study's results. In Hacking's (1965) words, probabilities are for "before trial betting," while likelihoods are for "after trial evaluation." An example can make the distinction more tangible: Suppose we are going to observe normal random variables with unit standard deviation, and we are interested in the mean, θ. We are particularly interested in the two hypotheses H_1: $\theta = 0$ and H_2: $\theta = 1$. Since the difference between these means is one standard deviation ($\Delta = 1$), figure 5.2 applies. Figure 5.2 shows that if we want to ensure that the probability of observing weak evidence with respect to these two hypotheses is less than .05, we must make at least $n = 18$ observations, and that with this sample size the probability of our observing misleading evidence (strong evidence in favor of one hypothesis when the other is true) is only $M(18) = .0045$.

I carried out such a study, generating 18 normal random variables with a common mean θ and unit standard deviation. The sample mean was 0.233, and the likelihood function, $L(\theta) = \exp[-18(0.233 - \theta)^2/2]$ (shown in figure 5.3), represents the evidence about the value of θ. These observations represent strong evidence supporting H_1 over H_2 ($L(0)/L(1) = 120$).

The fact that we considered only two specified values of θ when we planned the study does not preclude our examining the evidence as it relates to others, and figure 5.3 shows that some intermediate values are better supported than either $\theta = 0$ or $\theta = 1$, but the evidence supporting those values over $\theta = 0$ is weak. Nor does the fact that we planned to make only $n = 18$ observations preclude our deciding to make some more.

I made 4 more observations, looking at the evidence (the likelihood function) after each one. At that point I decided to double the sample size originally planned, and made 14 more observations. Figure 5.4 shows the likelihood function for all 36 observations. The evidence in favor of H_1 vs. H_2 is still strong ($L(0)/L(1) = 120$).

Finally, just to see what would happen, I decided to increase the sample

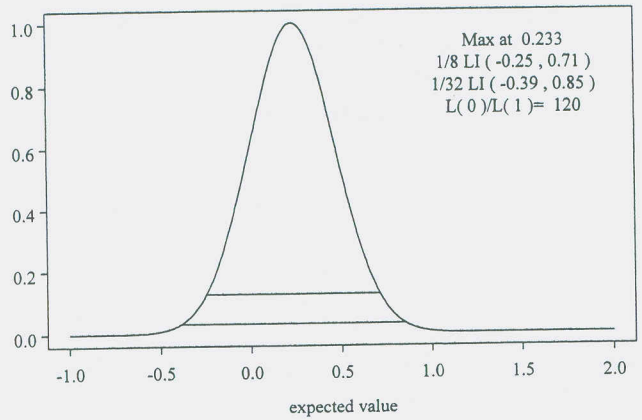

FIGURE 5.3 Likelihood for normal mean, $n = 18$ observations.

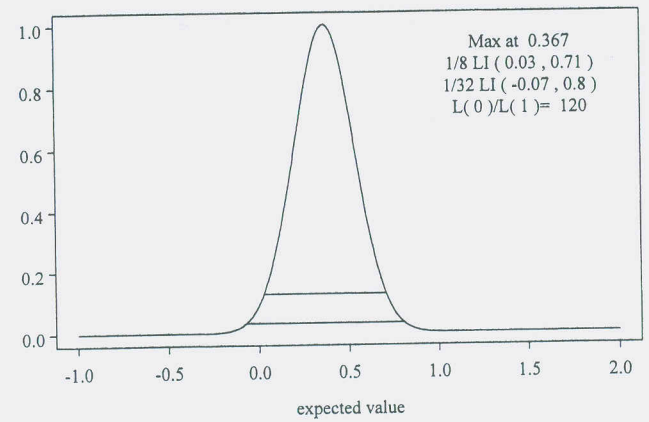

FIGURE 5.4 Likelihood for normal mean, $n = 36$ observations.

to $n = 100$. The evidence supporting $\theta = 0$ over $\theta = 1$ is now overwhelming, but figure 5.5 shows that we have strong evidence supporting values near 1/4 over $\theta = 0$.

What these 100 observations say about the value of θ is shown in figure 5.5. The probabilities in figure 5.2, which I used in choosing the initial sample size, are not relevant to the interpretation of this evidence. (For more on this point see Royall, 1997, sec. 4.5; 2000, rejoinder.) Nor is the fact that I considered only the two values $\theta = 0$ and $\theta = 1$ at the planning stage. The

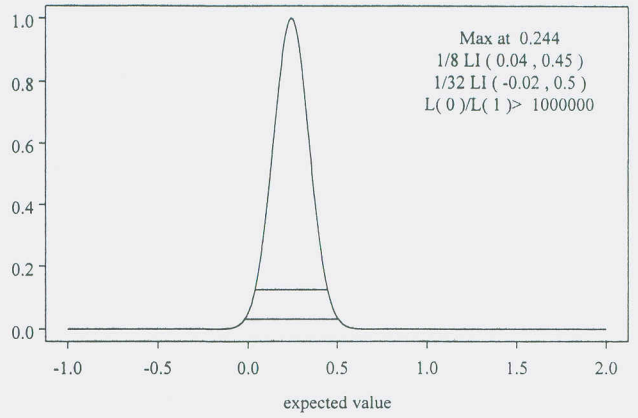

FIGURE 5.5 Likelihood for normal mean, $n = 100$ observations.

same is true of the stopping rule—whatever these observations mean, as evidence about their expected value θ, it is independent of the stopping rule. This evidence is not weakened (or otherwise affected) by the fact that I "peeked" at the data along the way, or that my original target likelihood ratio was only 8. Such facts are critical for determining confidence coefficients, P-values, Neyman-Pearson error probabilities (α, β), etc., but they have no effect on the likelihood function and no valid role to play in interpreting these 100 observations as evidence about θ.[2]

CONCLUSION

In a critique of Neyman-Pearson theory John Pratt (1961) rightly observed, "of course, skillful people can do useful statistics using Neyman and Pearson's formulation. But so can they using Fisher's or Jeffreys', or minimax decision theory, or subjective probability and Bayes' Theorem." Nevertheless, when the purpose of a statistical analysis is to represent and interpret data as evidence, all of these methods have serious shortcomings. These can be overcome within the frequentist approach by making explicit what has heretofore been treated only implicitly and intuitively—the concept of statistical

2. These observations are actually 100 random normal deviates with mean $\theta = 0.3$, and standard deviation 1.

evidence. The concept embodied in the law of likelihood, and represented in the terms "weak evidence" and "misleading evidence" that are central to the likelihood "evidence-generating" paradigm, can lead to a body of statistical theory and methods that:

1. requires probability models for the observable random variables only (and is in that sense frequentist, not Bayesian);
2. contains a valid, explicit, objective measure of the strength of statistical evidence;
3. provides for explicit, objective measure (and control) of the probabilities of observing weak or misleading evidence.

Under the likelihood paradigm, changing the stopping rule changes the probabilities of observing weak or misleading evidence, but these probabilities are used for planning only and do not affect the objective assessment of the observations as statistical evidence. There is no sound statistical reason to stop a study when the evidence is weak—it is quite appropriate to choose an initial sample size (perhaps on the basis of the *probability* of finding strong evidence, as in figure 5.2), evaluate the evidence in the observed sample, and then decide whether or not to make more observations. The scientist who carefully examines and interprets his observations (via likelihood functions) as they arise is behaving appropriately. He is not weakening or damaging the statistical evidence, he is not spending or using up statistical capital, he should not be chastized for peeking at the data, and there is no valid basis for levying a statistical fine on his study.

5.1 Commentary

D. R. Cox

First, it should be stressed that the law of likelihood is a very appealing working hypothesis, but in its strong sense it is no more than that. From several points of view two sets of data having equivalent likelihood functions should be regarded as providing the same evidence so long as the probability model from which they are derived is a secure base for interpretation. It is not at all so obvious that the same is true when the likelihoods come from different probability models. Indeed, it is a basic precept of the design of experiments and of survey sampling that the way the data are obtained should

be taken account of in analysis and in that sense dependence on a stopping rule, although a nuisance, is to be expected. Of course in many, but not all, contexts the dependence is so slight that, however interesting its presence or absence from a conceptual point of view, its practical relevance when treated correctly is negligible.

It is entirely sensible to begin the discussion of fundamental issues by looking at a very idealized situation, one with no nuisance parameters and just two possible hypotheses. Yet to some extent an approach has to be judged by the range of issues to which it will contribute. In some special cases the presence of nuisance parameters can be evaded by suitable modifications of the likelihood function, obtained, for example, by conditioning or marginalizing, but to be valuable an approach needs to go far beyond that. In some ways more serious is the matter of questions made precise only in the light of the data. For example, a particular value of a parameter may be compared with the maximum likelihood estimate. It is not at all clear that a given numerical value of the likelihood ratio has the same interpretation as in the simpler case. In both cases further calibration of likelihood-based statistics seems needed, and it is of course at this point that Richard Royall and I part company.

The discussion of the choice of sample size is illuminating, although for many purposes a formulation in terms of estimation is more appealing. That is, the objective is regarded as estimation of the parameter with a concentration of the likelihood function that depends in an approximately preassigned way on the location of the likelihood. Then some difficulties disappear. The final comment in the paper justifying continuous inspection of the data when data accrue sequentially also has much appeal. Indeed, I believe that many statisticians approaching statistics from a broadly frequentist perspective are uneasy at notions such as "spending error rates," perhaps because these treat notions of error rate as more than just hypothetical concepts used for calibrating measures of uncertainty against performance in idealized situations. While in some situations there may be compelling quasi-political arguments, as well as cost considerations, pointing against too frequent an analysis, in principle it is hard to see an argument at a completely fundamental level.

A full discussion of the evaluation of evidence needs to include procedures for checking the adequacy of any proposed model. These can be forced into a likelihood formulation only in a very restrictive sense.

Two further references that readers may find helpful are Barnard (1947) and Birnbaum (1969). Barnard was the first, I believe, to give the interesting inequality relating error rates to likelihood ratio. Birnbaum describes in

more detail why after much careful thought he had come to the conclusion that, the likelihood principle not withstanding, "sensible" confidence intervals were the most fruitful approach to the evaluation of evidence.

5.2 Commentary

Martin Curd

Professor Royall's paper is both elegant and challenging. In my remarks, I focus on the contrast between the likelihood conception of evidence and the Bayesian account. I do this because, despite the rarity of the Bayesian approach among professional statisticians (Moore, 1997), it has been a dominant influence on recent attempts by philosophers to understand the confirmation of scientific theories (Horwich, 1982; Salmon, 1990; Earman, 1992; Howson and Urbach, 1993). My main concern is whether the likelihood paradigm of evidence can be plausibly extended from the realm of statistics to scientific theories in general.

According to the standard Bayesian approach, a hypothesis H is confirmed by an observation E (and thus E is evidence for H) if and only if the posterior probability $P(H|E)$ is greater than the prior probability $P(H)$. The two are related by the Bayesian equation $P(H|E) = [P(H) \times P(E|H)]/P(E)$, where $P(E) > 0$. $P(E|H)$ is the likelihood of H, and $P(E)$ the expectedness of the evidence. (To avoid clutter, the reference to background knowledge in each of the factors in Bayes' equation has been omitted.) Although other measures have been proposed, Bayesians often take the strength of E as evidence for H (the degree of confirmation that E confers on H) to be some positive increasing function of $P(H|E) - P(H)$, the difference between the posterior probability of H and its prior probability (Howson and Urbach, 1993; Christensen, 1999). Of the three terms on the right-hand side of Bayes' equation, the likelihood $P(E|H)$ is the least controversial since scientific theories are often formulated such that (in conjunction with appropriate auxiliary hypotheses and background assumptions) we can deduce from them the probability of events such as E. Thus, $P(E|H)$ can be regarded as being objective. But what about $P(H)$ and $P(E)$? Some Bayesians are unashamedly subjectivist about $P(H)$ and regard the prior probability of H as simply the degree of belief that a particular scientist happens to have in H. Others have tempered this subjectivism by placing constraints on permissible priors (for example, no one is allowed to assign a prior of zero to any noncontradictory

hypothesis, since this would prevent that hypothesis from ever being confirmed) and by regarding $P(H)$ as our informed judgment about the plausibility of H in light of such factors as simplicity, explanatory power, and the track record of similar hypotheses in the past. Bayesians also appeal to the "swamping" argument in mitigation of the charge of subjectivism. For, as each new piece of evidence E_i is considered, Bayesians replace the "old prior" $P(H)$ with the "new prior" $P(H|E_i)$, calculated using Bayes' equation. Although scientists may initially assign widely different (nonzero) priors to H, the values of $P(H|E_i)$ will converge as evidence accumulates (as long as the scientists agree on the likelihoods). Thus, subjective disagreements about the plausibility of hypotheses (reflected in different initial assignments of priors) have a diminishing influence on the confirming power of new evidence. In the long run, the initial priors become irrelevant.

Much ink has been spilled on the difficulty of interpreting $P(E)$ since Glymour (1980) first drew attention to the problem of old evidence; that is, the problem of assigning to $P(E)$ a value less than 1 when E becomes known and thus presumably certain for the scientist who knows it. One approach is to expand $P(E)$ as a sum of terms of the form $P(H_i) \times P(E|H_i)$, where the set $\{H_i\}$ includes H and all the logically possible alternative hypotheses that also predict E. But in practice, the number of alternatives that have been formulated is small, and so typically we have to invoke the notorious catchall hypothesis, H_c, to complete the set. Unfortunately, the catchall is not a definite hypothesis at all but says merely "none of the above"; and so we cannot use it to calculate $P(E|H_c)$ and thus complete the calculation of $P(E)$. Another approach (Howson and Urbach, 1993) is to say that $P(E)$ is the probability that would be assigned to E independently of consideration of H. We are enjoined to imagine that we delete H (and anything that entails H) from our corpus of belief and then see what probability would be assigned to E given our background knowledge. While the details of this deletion procedure are hard to make precise, there are cases in which our background information can provide a reasonable basis for estimating $P(E)$. (See the discussion of the raven paradox at the end of this commentary.)

Royall thinks that the standard Bayesian account of evidence is fundamentally flawed because of its reliance on subjective probabilities such as $P(H)$. He argues that while Bayes' equation describes how one should change one's beliefs in response to new information, it is a mistake to think that any function of prior and posterior probabilities could measure (objective) evidential strength. Rather, for rival hypotheses H and J, that role is played by the likelihood ratio, $P(E|H)/P(E|J)$. As stated by Royall, the law of likelihood says two things. First, it gives a sufficient condition for an event, E, to con-

fer greater evidential support on hypothesis H than on hypothesis J. It says that if $P(E|H) > P(E|J)$, then E supports H more than it does J. The second part says that the likelihood ratio measures the strength of that comparative support.

A striking feature of the likelihood paradigm is its insistence that evidential support is essentially comparative. If Royall is right, there is literally no such thing as the support that a prediction E gives to a hypothesis H; there is only the degree of support that E gives to H as compared with a rival hypothesis, J. Note that there is an important difference between evidential support being by its very nature comparative and its being contextual. Bayesians agree that evidential support is contextual since the derivation of predictions from H and the estimation of $P(E)$ will depend on other hypotheses and background information. But although the Bayesian algorithm requires contextual input, its output (telling us the degree to which E supports H) is noncomparative: it is not limited, in principle, to comparing E's support of H with E's support of J. That we can make comparative judgments in contexts where absolute measures are difficult or practically impossible should occasion no surprise. For example, we can judge which of two sticks is the longer or which of two objects is the heavier without having first to ascertain the length or weight of each. But in these cases, underlying a comparative measure there is a pair of absolute values that grounds the comparison. Why, according to the likelihood paradigm, should statistical evidence be so different in this regard? How, one might naïvely ask, can we make sense of the notion that E supports H more than it does J if there is no such thing as E's support for H or E's support for J? Moreover, if we try to generalize the likelihood conception of comparative evidence into a general account of evidence for scientific theories, its scope (unlike that of the Bayesian account) would be severely limited. For it would apply only to the comparison of pairs of rival theories with respect to the same evidence, leaving unanswered such questions as "Is H supported more strongly by evidence E than J is by evidence F?"

As stated by Royall, the law of likelihood gives a sufficient condition for greater evidential support, not a necessary condition, and so it implies nothing about cases in which the likelihoods of competing hypotheses are the same. But it would appear to be in the spirit of the likelihood account that, when the likelihoods $P(E|H)$ and $P(E|J)$ are equal, neither H nor J receive greater support from E. Either each receives no support from E or they each receive the same degree of support. As far as E is concerned, H and J are evidentially on a par. Similarly, for separate pieces of evidence E and F, if $P(E|H)$ and $P(F|H)$ are the same, so is their confirming power for H. But this

consequence of the likelihood paradigm—same likelihood, same confirming power—seems contrary to our intuitions about evidence. To illustrate the problem for the likelihood account, consider two cases: the problem of irrelevant conjunction and the raven paradox.

The problem of irrelevant conjunction arises when we have two hypotheses, H and $(H \& I)$, where I is a contingent statement that is logically independent of H and irrelevant to E. For simplicity's sake, I shall focus on the case in which H is deterministic, but the same problem can arise when H is statistical. H (in conjunction with background information) entails a true observational prediction, E. It follows that the augmented theory $(H \& I)$ also makes the same prediction, and the two likelihoods, $P(E|H)$ and $P(E|H \& I)$, are each equal to 1. Nonetheless, it is widely held that E confirms $(H \& I)$ less strongly than it does H. For example, the observation of a white swan provides weaker support for "All swans are white and some gazelles are brown" than it does for "All swans are white." The Bayesian account can do justice to this intuition by appealing to the role of prior probabilities in confirmation. For on the Bayesian analysis we have $P(H|E) - P(H) = P(H) \times [(1 - P(E))/P(E)]$, and $P(H \& I|E) - P(H \& I) = P(H \& I) \times [(1 - P(E))/P(E)]$. The factor in the square brackets is the same for both theories, and so their respective degrees of confirmation are proportional to the prior probabilities, $P(H \& I)$ and $P(H)$. Since $(H \& I)$ entails H, and I is a contingent statement that is independent of H, $P(H \& I)$ must be less than $P(H)$. Thus, E confirms $(H \& I)$ by a smaller amount than it confirms H; adding the irrelevant conjunct to H lowers the confirmation provided by E. In this respect, then, the Bayesian approach to confirmation appears to be superior to the likelihood account.

The key to the Bayesian solution of the raven paradox lies in the different values assigned to $P(E)$ by our background knowledge. Let H be the hypothesis "All ravens are black," which we shall write as "All Rs are B." Since H is logically equivalent to its contrapositive, "All non-Bs are non-Rs," it would seem that H should be confirmed not only by the observation of a black raven but also by the observation of a nonblack nonraven (such as a white shoe). Bayesians see no paradox here. They argue that the observation of a white shoe does confirm H, but only to a very small degree, a degree that is much smaller than the confirmation conferred on H by the observation of a black raven. In this way, Bayesians explain why many people regard the raven case as paradoxical when they first encounter it; for it is understandable that most of us are unable to distinguish a very low degree of confirmation from no confirmation at all.

Following Horwich (1982) we adopt a notation that reflects the manner

in which evidence is collected: $(R * B)$ is the discovery that a randomly selected object that is already known to be a raven is black; $(\sim B * \sim R)$ is the discovery that a randomly selected object that is already known to be nonblack is a nonraven. The asterisk indicates which component of each paired observation is made first. To repeat, the observation reports include information about the method used to generate the report. That said, it must be the case that both likelihoods, $P(R * B|H)$ and $P(\sim B * \sim R|H)$, are equal to 1; for, if all ravens are black, then the probability that a raven will turn out to be black is 1; similarly, given H, the probability that a nonblack thing will turn out to a nonraven is also 1. Thus, the Bayesian comparison of the confirming power of the observation report $(R * B)$ with the confirming power of the observation report $(\sim B * \sim R)$ depends on the inverse ratio of their probabilities. In order to estimate those probabilities, we need to specify some background information. Let x be the fraction of things in the universe that are ravens, let y be the fraction of things that are black, and let α be the fraction of things that initially are believed to be black ravens. This yields $P(R * B) = \alpha/x$, and $P(\sim B * \sim R) = [(1 - y) - (x - \alpha)]/(1 - y)$. Comparison of these two expressions shows that $(R * B)$ must support H more strongly than does $(\sim B * \sim R)$ if x is greater than α and $(1 - y)$ is greater than x. Thus, if we do not already believe that all ravens are black, $(R * B)$ is stronger evidence for H than is $(\sim B * \sim R)$ as long as we also believe that nonblack things are more abundant than ravens. Both sorts of observation should increase our confidence in H, but finding black ravens provides stronger support.

It is instructive to compare this Bayesian treatment of the raven paradox with the likelihood analysis in Royall (1997). On Royall's likelihood analysis, the likelihoods of $(R * B)$ and $(\sim B * \sim R)$ are not both equal to 1. They differ because the likelihood of each observation report is assessed with respect to H as compared with the rival hypothesis J (according to which the proportion of ravens that are black is some fraction less than 1); and, in the case of $(\sim B * \sim R)$ reports, the calculation of the likelihood $P(\sim B * \sim R|J)$ depends on the same background information about the relative abundance of ravens and nonblack things as in the Bayesian analysis. Royall agrees that when the sampling procedure yields reports of the form $(\sim B * \sim R)$, the observation of a nonblack nonraven has the power to confirm H; and that when the number of nonblack things in the universe vastly exceeds the number of ravens, the confirming power of such observations is very weak. But these conclusions about evidential support, based on the likelihood paradigm, are essentially comparative: observations of the type $(\sim B * \sim R)$ provide only marginally stronger support for H than for J. Similarly, observations of the type $(R * B)$ support H more strongly than they do J. Unlike the

Bayesian account, the likelihood analysis yields no conclusion about the relative confirming power for H of the two different types of observation. Because it is limited to judging the relative support given to rival hypotheses by the same observation report, the likelihood analysis leaves unanswered the crucial question that lies at the heart of the raven paradox: why do reports of the form $(R * B)$ confirm H much more strongly than do reports of the form $(\sim B * \sim R)$?

5.3 Rejoinder

Richard Royall

Professor Cox rejects the law of likelihood out of hand, expressing the opinion that it represents nothing more profound than a "working hypothesis." He goes on to describe some aspects of my paper that he finds appealing and some that he does not.

It is disappointing that Cox chooses to reveal so little about the rationale for his judgments. For instance, he apparently embraces a "basic precept of the design of experiments . . . that the way the data are obtained should be taken account of in analysis." In my opinion it would be quite useful to the discipline of statistics if Cox would formalize and elaborate on this "basic precept"[3] in such a way that its implications could be seen (and its validity tested) in specific examples such as the one I use to illustrate the irrelevance of stopping rules in analyses *whose purpose is to interpret observed data as evidence.* Does his precept imply that the meaning of the coin-toss observations that you and I made together, (1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1), as evidence about the tendency of the 40¢ coin to fall heads depends on whether we stopped the experiment because we had made 17 tosses (my preferred stopping rule), or because we had seen 9 heads (your rule)? . . . or because the coin fell apart? . . . or because it was time for tea? If his precept does indeed have such implications then it would suggest to me that Cox should correct the precept, not reject the law of likelihood.

In the meantime, statistics remains in a theoretical and conceptual mess —a "synthesis of ingredients borrowed from mutually incompatible theoretical sources" (Birnbaum, 1970, 1033). With no accepted principles to

3. In Royall (1991), I tried to formalize and criticize such a precept, which I called the "randomization principle."

guide statistical reasoning, we can offer scientists who are perplexed by our controversies (such as the one about "spending error rates") nothing more than conflicting expert judgments about what is sensible. I cannot share Professor Cox's satisfaction with this state of our discipline.

I now reply to comments by Martin Curd. The law of likelihood answers a fundamental question about empirical evidence: When does an observation constitute evidence supporting one hypothesis vis-à-vis another? The law says it is when the two hypotheses imply different probabilities for the observation. It says that the hypothesis that implies the greater probability is the better supported and that the probability ratio measures the strength of the evidence.

Professor Curd contrasts this concept of evidence to the Bayesian account, where only one hypothesis is made explicit, the question being, When does an observation constitute evidence supporting a hypothesis? The Bayesian answer is, When the observation has the effect of increasing the probability of the hypothesis. Preferring the Bayesian account, Professor Curd challenges the law of likelihood by presenting two examples where it purportedly leads to conclusions that seem "contrary to our intuitions about evidence."

I want to suggest that the law of likelihood is not a threat, or even an alternative, to the Bayesian view, but that to the contrary it constitutes the essential core of the Bayesian account of evidence. My claim is that the Bayesian who rejects the law of likelihood undermines his own position. Then I will argue that this act of self-destruction is unwarranted, because Professor Curd's argument leading to rejection of the law springs from a misunderstanding. In this discussion I will assume that the probabilities obey the usual (Kolmogorov) axioms. Since Bayes' theorem is derived from those axioms, I find it hard to imagine a Bayesian who rejects them.

According to Professor Curd, the Bayesian interpretation is that an observation E as evidence for H when $P(H|E) > P(H)$. His analysis assumes the existence of the three terms, $P(E|H)$, $P(H)$, and $P(E)$, that appear on the right-hand side of what he calls Bayes' equation,

$$P(H|E) = P(E|H)P(H)/P(E).$$

Now if H has probability $P(H)$ then the axioms imply that its negation $\sim H$ has probability $1 - P(H)$, and that furthermore $P(E)$ can be expressed as[4]

4. This result is a special case of what is sometimes called the theorem on total probability, which is essential to Bayes' theorem (Kolmogorov, 1956, sec. 4).

$$P(E) = P(E|H)P(H) + P(E|\sim H)[1 - P(H)].$$

Rearranging this expression, we see from the three quantities that the Bayesian must supply, $P(E|H)$, $P(H)$, and $P(E)$, we can deduce the value for the probability of E that is implied by the hypothesis[5] $\sim H$:

$$P(E|\sim H) = [P(E) - P(E|H)P(H)]/[1 - P(H)].$$

If we return to the critical inequality, $P(H|E) > P(H)$, and replace $P(H|E)$ with the expression given by Bayes' equation, $P(H|E) = P(E|H)P(H)/P(E)$, we see that the Bayesian interprets E as evidence for H if and only if $P(E|H) > P(E)$. And if in this last inequality we substitute the expression for $P(E)$ displayed above, we see that the Bayesian's criterion is equivalent to $P(E|H) > P(E|H)P(H) + P(E|\sim H)[1 - P(H)]$, which is equivalent to

$$P(E|H) > P(E|\sim H).$$

That is, $P(H|E) > P(H)$ if and only if $P(E|H) > P(E|\sim H)$.

When is E "evidence for H" in the Bayesian scheme? When does $P(H|E)$ exceed $P(H)$? It is precisely when H implies a greater probability for E than the alternative hypothesis $\sim H$ does. It is precisely when the law of likelihood says that E is evidence supporting H over $\sim H$.

The Bayesian may prefer not to make $P(E|\sim H)$ explicit—he may prefer to leave this object buried within $P(E)$. But whether he chooses to make it explicit or not, the value of $P(E|\sim H)$ is determined by $P(E|H)$, $P(E)$, and $P(H)$. And when all the cards are laid on the table, the winner of the Bayesian's game is decided according to the law of likelihood—the hypothesis (H or $\sim H$) that is better supported by the observation E is the one that implies the greater probability for that observation. *Hypothesis H "is confirmed by" E if and only if E is more probable if H is true than if H is false.* The Bayesian's qualitative conclusion (e.g., that E is evidence *for* H rather than against it) must conform to the law of likelihood.

5. The same deviation applies when there is a set of disjoint hypotheses $\{H_i\}$. If $\sum P(H_i) < 1$, then from $P(H_i)$, $P(E|H_i)$, and $P(E)$, we can deduce the probability of E under the "notorious catchall hypothesis, H_c," i.e., $P(E|H_c) = [P(E) - \sum P(E|H_i)P(H_i)]/[1 - \sum P(H_i)]$. The Bayesian cannot simultaneously claim (i) to know the terms on the right-hand side of this equation and (ii) that $P(E|H_c)$ is nonexistent or unknowable without violating the axioms of probability theory.

But are the two accounts of evidence in *quantitative* accord? One says unequivocally that the strength of the evidence for H versus $\sim H$ is measured by the likelihood ratio $P(E|H)/P(E|\sim H)$.[6] The other, according to Professor Curd, is less definite: "Bayesians often take the strength of E as evidence for H to be some positive increasing function of $P(H|E) - P(H)$." As we have seen, the difference $P(H|E) - P(H)$ does point in the right direction. But as a measure of the *strength* of the evidence it is curious. It says that E cannot be strong evidence for H when $P(H)$ is large, i.e., when there is strong prior evidence for H. Thus E can be strong evidence for H when it flies in the face of extensive previous experience ($P(H)$ is small), but not when it is consistent with that experience ($P(H)$ near 1). Now, it is plausible that E is, in some sense, more valuable, more important, or more newsworthy when it elevates a previously implausible hypothesis to respectability than when it merely confirms what was already believed. But that E is *stronger evidence* in the former case is not at all clear.

This counterintuitive aspect of the way the difference, $P(H|E) - P(H)$, depends on the prior probability $P(H)$ suggests that the Bayesian should adopt some other measure. In fact the Bayesian I. J. Good (1968) proposed some desiderata for an evidence measure and proved they imply that the measure should be an increasing function (the logarithm) of the likelihood ratio $P(E|H)/P(E|\sim H)$, which is independent of $P(H)$.

Thus, it is clear to me that the Bayesian cannot escape either the qualitative or quantitative conclusions of the law of likelihood. Professor Curd, on the other hand, sees a problem with "the likelihood account" of evidence, and he gives two examples intended to illustrate that problem. What I think they actually illustrate is the misconception that is expressed in his statement of the problem: "for separate pieces of evidence E and F, if $P(E|H)$ and $P(F|H)$ are the same, so is their confirming power for H. But this consequence of the likelihood paradigm—same likelihood, same confirming power—seems contrary to our intuitions about evidence."

What *does* the law of likelihood say when one hypothesis attaches the same probability to two different observations? It says absolutely nothing. The law of likelihood applies when two different hypotheses attach probabilities to the same observation. It states that the ratio of the probabilities, $P(E|H)/P(E|J)$, measures the evidence in the observation E for hypothesis H vis-à-vis hypothesis J. Although it says that their ratio measures the evidence, the law attaches no evidential meaning to either probability in isolation. In the absence of an alternative hypothesis, $P(E|H)$ is not the support for H, the strength of the evidence for H, or the confirming power of E for H.

6. The answer to Professor Curd's question, "Is H supported more strongly by evidence E than J is by evidence F?" is straightforward: E supports H versus $\sim H$ more strongly than F supports J versus $\sim J$ if and only if $P(E|H)/P(E|\sim H) > P(F|J)/P(F|\sim J)$.

Consider an example: Let H be the hypothesis that the proportion of white balls in an urn is 1/2. For the two observations

E: 1 white ball in 1 draw and
F: 10 or fewer white balls in 21 independent draws,

the probabilities $P(E|H)$ and $P(F|H)$ are the same—both equal 1/2. But the law of likelihood says nothing about the "confirming power" of E for H or of F for H. It does address the interpretation of these observations as evidence for H vis-à-vis any other hypothesis that also implies probabilities for them, such as the hypothesis J that the proportion of white balls is 1/4. Since $P(E|J) = .25$ and $P(F|J) = .994$, the likelihood ratios are $P(E|H)/P(E|J) = 2.0$ and $P(F|J)/P(F|H) = 1.987$, so the law of likelihood says that E supports H over J, that F supports J over H, and that the evidence has nearly the same (weak) strength in both cases.

Now consider a different alternative to H. Let K be the hypothesis that the proportion of white balls is 3/4 so that $P(E|K) = .75$ and $P(F|K) = .0064$. Observation E is weak evidence supporting K over H (likelihood ratio = 1.5), while F is strong evidence for H versus K (likelihood ratio = 78). Thus, although observations E and F are equally probable under hypothesis H, E is evidence *for* H vis-à-vis hypothesis J, and *against* H vis-à-vis K. And in each case the observation F has the opposite evidential meaning.

The likelihood view is that observations like E and F have no valid interpretation as evidence in relation to the single hypothesis H. I have discussed elsewhere (Royall, 1997, sec. 3.3) the futility of efforts to attach evidential meaning to an observation in relation to a single hypothesis that is not logically incompatible with the observation.

Finally, I want to compare the Bayesian and likelihood analyses of the raven paradox. Professor Curd discusses the paradox in terms of the proportions in a 2×2 table:

	black	nonblack	
ravens	α	$x - \alpha$	x
nonravens	$y - \alpha$	$1 - x - y + \alpha$	$1 - x$
	y	$1 - y$	1

Here x represents the proportion of the objects in the population under consideration that are ravens, y is the proportion of objects that are black, and α is the proportion that are black ravens.

An observation $R * B$ represents an object drawn at random from the first row that is found to come from the first column (a raven that is found to be black), and an observation $\sim B * \sim R$ represents a draw from the second *column* that is found to come from the second row (a nonblack thing that proves to be a nonraven). The likelihood analysis is as follows:

> The hypothesis H (all ravens are black) asserts that α equals x, implying that the probability of $R * B$ is 1.
> The alternative hypothesis, J, asserts that α has a value less than x, which implies that the probability of $R * B$ is $\alpha/x < 1$.
> Therefore the law of likelihood says the observation $R * B$ is evidence supporting H over J by the factor: $P(R * B|H)/P(R * B|J) = 1/(\alpha/x) = x/\alpha$.

Similarly H implies that the probability of $\sim B * \sim R$ is 1, while J implies that the probability is $(1 - x - y + \alpha)/(1 - y) < 1$, so the law of likelihood says the observation $\sim B * \sim R$ is evidence supporting H over J by

$$\frac{P(\sim B * \sim R|H)}{P(\sim B * \sim R|J)} = \frac{1}{(1 - x - y + \alpha)/(1 - y)} = \frac{1 - y}{1 - x - y + \alpha}.$$

These are exactly the same expressions on which Professor Curd bases his Bayesian analysis. His measures of "the relative confirming power for H of the two different types of observation" are just the two likelihood ratios.

As I noted earlier, the Bayesian may choose not to speak of the alternative to H. His use of the expression "the relative confirming power for H of the two different types of observation," without reference to an alternative, obscures the fact (revealed in the appearance of the quantity α in his formulae) that the observation can properly be said to confirm H only in relation to an alternative hypothesis that assigns lower probability to that observation than H does. It obscures the fact that his statement "reports of the form $(R * B)$ confirm H much more strongly than do reports of the form $(\sim B * \sim R)$" means nothing more or less than that the first report's likelihood ratio, $P(R * B|H)/P(R * B|J) = x/\alpha$, is the greater.

The Bayesian's failure to make explicit the alternative hypothesis that is implicit in his analysis obscures the fact that the rock on which his analysis is built is the law of likelihood.

REFERENCES

Barnard, G. A. 1947. Review of *Sequential Analysis*, by A. Wald. *J. Am. Stat. Assn* 42: 658–664.

Berger, J. O., and Bernardo, J. M. 1992. On the Development of Reference Priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4*. New York: Oxford University Press.

Berger, J. O., and Wolpert, R. L. 1988. *The Likelihood Principle*. 2nd ed. Hayward, CA: Institute of Mathematical Statistics.,

Bernardo, J. M. 1979. Reference Posterior Distributions for Bayesian Inference (with discussion). *J. Roy. Stat. Soc.*, ser. B, 41:113–147.

Birnbaum, A. 1962. On the Foundations of Statistical Inference (with discussion). *J. Am. Stat. Assn* 57:269–326.

Birnbaum, A. 1969. Concepts of Statistical Evidence. In. Morgenbesser, S., P. Suppes, and M. White, eds., *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. New York: St. Martin's Press.

Birnbaum, A. 1970. Statistical Methods in Scientific Inference. *Nature* 225:1033.

Bower, B. 1997. Null Science: Psychology's Status Quo Draws Fire. *Science News* 151 : 356–357.

Christensen, D. 1999. Measuring Confirmation. *J. Phil.* 96:437–461.

Cohen, J. 1994. The Earth Is Round ($p < .05$). *Am. Psych.* 49:997–1003.

Earman, J. 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge: MIT Press.

Edwards, A. W. F. 1969. Statistical Methods in Scientific Inference. *Nature* 222: 1233–1237.

Edwards, A. W. F. 1972. *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*. Cambridge: Cambridge University Press.

Edwards, A. W. F. 1992. *Likelihood* (expanded ed.). Baltimore: Johns Hopkins University Press.

Edwards, W., H. Lindman, and L. Savage. 1984. Bayesian Statistical Inference for Psychological Research. In Kadane, J. B., ed., *Studies in Bayesian Econometrics*. (Originally published in *Psych. Rev.* 70:193–242.)

Efron, B. 1986. Why Isn't Everyone a Bayesian? (with discussion), *Am. Stat.* 40:1–11.

Fisher, R. A. 1959. *Statistical Methods and Scientific Inference*. 2nd ed. New York: Hafner.

Glymour, C. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Good, I. J. 1968. Corroboration, Explanation, Evolving Probability, Simplicity, and a Sharpened Razor. *Br. J. Phil. Sci.* 19:123–143.

Goodman, S. N. 1998. Multiple Comparisons, Explained. *Am. J. Epidemiol.* 147: 807–812.
Hacking, I. 1965. *Logic of Statistical Inference,* New York: Cambridge University Press.
Horwich, P. 1982. *Probability and Evidence.* Cambridge: University of Cambridge Press.
Howson, C., and Urbach, P. 1993. *Scientific Reasoning: The Bayesian Approach.* 2nd ed. Peru, IL: Open Court.
Jeffreys, H. 1961. *Theory of Probability.* 3rd ed. Oxford: Oxford University Press.
Kass, R. E., and A. E. Raftery. 1995. Bayes Factors. *J. Am. Stat. Assn* 90:773–795.
Kolmogorov, A. N. 1956. *Foundations of Probability.* Trans. N. Morrison. New York: Chelsea.
Lindley, D. V. 1965. *Introduction to Probability and Statistics: Part I,* Cambridge: Cambridge University Press.
Lindley, D. V. 1992. Discussion of Royall, R. M.: The Elusive Concept of Statistical Evidence. In Bernardo, J. M., J. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4.* New York: Oxford University Press.
Moore, D. S. 1997. Bayes for Beginners? Some Reasons to Hesitate. *Am. Stat.* 51: 254–261.
Morrison, D. E., and Henkel, R. E. 1970. *The Significance Test Controversy.* Chicago: Aldine.
Neyman, J. 1950. *First Course in Probability and Statistics.* New York: Henry Holt.
Pagano, M., and K. Gauvreau. 1993. *Principles of Biostatistics,* Belmont, CA: Duxbury,
Pratt, J. W. 1961. Review of *Testing Statistical Hypotheses,* by E. L. Lehmann (1959). *J. Am. Stat. Assn* 56:163–166.
Robbins, H. 1970. Statistical Methods Related to the Law of the Iterated Logarithm. *Ann. Math. Stat.* 41:1397–1409.
Royall, R. M. 1991. Ethics and Statistics in Randomized Clinical Trials (with discussion). *Stat. Sci.* 6:52–88.
Royall, R. M. 1997. *Statistical Evidence: A Likelihood Paradigm.* London: Chapman and Hall.
Royall, R. M. 2000. On the Probability of Observing Misleading Statistical Evidence (with discussion). *J. Am. Stat. Assn* 95:760–780.
Salmon, W. C. 1990. Rationality and Objectivity in Science, or Tom Kuhn Meets Tom Bayes. In Savage, C. W., ed., *Scientific Theories.* Minneapolis: University of Minnesota Press.
Savage, L. J. 1962. Discussion of A. Birnbaum, On the Foundations of Statistical Inference. *J. Am. Stat. Assn* 53:307–308.
Smith, C. A. B. 1953. The Detection of Linkage in Human Genetics. *J. Roy. Stat. Soc.,* ser. B,15:153–192.
Sterne, J. A. C., and Smith, G. D. 2001. Sifting the Evidence: What's Wrong with Significance Tests? *BMJ* 322:226–231.
Thompson, J. R. 1998. Invited commentary on Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data. *Am. J. Epidemiol.* 147:801–806.

6 Why Likelihood?

Malcolm Forster and Elliott Sober

ABSTRACT

The likelihood principle has been defended on Bayesian grounds, on the grounds that it coincides with and systematizes intuitive judgments about example problems, and by appeal to the fact that it generalizes what is true when hypotheses have deductive consequences about observations. Here we divide the principle into two parts—one qualitative, the other quantitative—and evaluate each in the light of the Akaike information criterion (AIC). Both turn out to be correct in a special case (when the competing hypotheses have the same number of adjustable parameters), but not otherwise.

INTRODUCTION

Mark Antony said that he came to bury Caesar, not to praise him. In contrast, our goal is neither to bury the likelihood concept nor to praise it. Instead of praising it, we will present what we think is an important criticism. However, the upshot of this criticism is not that likelihood should be buried, but a justification of likelihood, properly understood.

Before we get to our criticism of likelihood, we should say that we agree with the criticisms that likelihoodists have made of Neyman-Pearson-Fisher statistics and of Bayesianism (Edwards, 1987; Royall, 1997). In our opinion, likelihood looks very good indeed when it is compared with these alternatives. However, the problem of a positive defense of likelihood remains. Royall begins his excellent book with three kinds of justification.

We would like to thank Ken Burnham, Ellery Eells, Branden Fitelson, Ilkka Kieseppä, Richard Royall, and the editors of this volume for helpful comments on an earlier draft.