LIKELIHOOD AND CONVERGENCE*

ELLIOTT SOBER†

Philosophy Department
University of Wisconsin, Madison

A common view among statisticians is that convergence (which statisticians call consistency) is a necessary property of an inference rule or estimator. In this paper, this view is challenged by appeal to an example in which a rule of inference has a likelihood rationale but is not convergent. The example helps clarify the significance of the likelihood concept in statistical inference.

1. Introduction. Maximum likelihood estimation (*MLE*) is a technique for estimating probabilities from sample frequencies. Suppose an urn is filled with balls that differ in color. We repeatedly draw a ball, note its color, and return it to the urn (which we then shake). We find in our sample of twenty balls that a given proportion, *p*, are green. What is the best guess as to the probability of drawing a green ball? If the draws are independent, then *MLE* implies that the best estimate is that the probability is *p*. This estimate makes the observations maximally probable and is thereby said to be best supported by, or to best explain, the observations.

So it is said. But *why* should the best estimate be one that maximizes the probability of the observations? A common answer appeals to the fact that *MLE* has the desirable "operating characteristic" of convergence (which statisticians call consistency). As we increase our sample size, it becomes more and more probable that our estimate will fall within any fixed interval of the true value; in the limit, the probability goes to unity that our estimate will differ at most infinitesimally from the truth. *MLE* makes sense, in part, because it has this asymptotic property.

Indeed, *MLE* is convergent in the urn example. But let us look more carefully at the "because" that is said to link the justification of *MLE* with the asymptotic property. Three claims must be distinguished: (i) convergence is a *desirable* property of an estimator or rule of inference; (ii)

*Received April 1987.

†My thanks to Carter Denniston, A. W. F. Edwards, Ellery Eells, Joe Felsenstein, Isaac Levi, Steven Spielman, and the anonymous referees of this journal for useful discussion and to the National Science Foundation for financial assistance.

Philosophy of Science, 55 (1988) pp. 228–237.

convergence *suffices* to justify an estimator or rule of inference; (iii) convergence is a *necessary* property of a reasonable estimator or rule of inference.

The first of these I will not dispute, provided an "all else being equal" is appended. The second is known to be false, in that other estimators besides the one sanctioned by *MLE* are convergent. For example, if the sampling problem concerned inferring the average height in a normal population from measurements in a sample, the sample mean would be the *MLE* estimate, even though both the mean and the mode are convergent estimators of the population average. This fact is standardly interpreted as showing that convergence is not enough. Further properties must be produced that show why the mean is "better". It is here that the ideas of efficiency and unbiasedness come into play.

This leaves the third thesis: is convergence *necessary* for an estimator to be used? Statisticians frequently say that it is. Fisher (1950, p. 11) says that inconsistent estimators are "outside the pale of decent usage". Neyman (1952, p. 188) cites this pronouncement and emphatically agrees. Kendall and Stuart (1973, p. 273) say that "it seems perfectly reasonable to require consistency". These authors hold that preferring the maximum likelihood estimate of a parameter, or the hypothesis that is most likely in the light of the data, makes sense only in those contexts in which maximum likelihood possesses the asymptotic property of consistency.

A quite different tradition sees likelihood as a "primitive postulate" that does not stand in need of repeated sampling justifications. The words are Fisher's (1938); they stand in stark contrast with the position that Fisher espoused in the passage referred to above. Edwards (1972, p. 100) sees the primitive postulate idea as Fisher's mature position; he argues that although a wide range of likelihood procedures are statistically consistent, their claim on our attention does not depend on their being consistent. Hacking (1965, pp. 184–185) also is skeptical of the importance of consistency; although the maximum likelihood estimators he considered are consistent, he wondered if consistency might not be a "magical property" of those estimators and not in itself a "criterion of excellence". According to this point of view, likelihood measures the degree to which a hypothesis is supported by the evidence, regardless of whether the use of likelihood insures convergence.

The dispute as to whether an estimator must be consistent also arises when the question is whether a test of a hypothesis must be consistent. Since these are different concepts, it is worth stating each clearly. In estimation, one is interested in discovering the value of some parameter θ. A statistic s_n, which is a function of the n observations in the sample, is a consistent estimate of θ if and only if, for any two positive numbers d and e, a number n_0 exists such that when n exceeds n_0

$$Pr(|s_n - \theta| < d) > 1 - e.$$

In the case of hypothesis testing, a test of hypothesis H_0 against an alternative hypothesis H_1 is consistent precisely when, if H_1 is true, the probability of rejecting H_0 tends to unity as the sample size goes to infinity. We might roughly summarize what consistent estimation has in common with consistent testing by saying that a consistent procedure is one in which the rejection of falsehood (a mistaken estimate, a mistaken hypothesis) is virtually certain if the data are infinitely large.

If the use of maximum likelihood in estimation or in hypothesis testing always possessed the property of consistency, this foundational difference would not make a difference in practice. Different researchers then might concur that maximum likelihood methods should be used, but would disagree about the rationale for using them. However, the debate is not quite so cut off from practice. Although sufficient conditions of considerable generality have been established for the consistency of maximum likelihood methods (Wald 1949), it has been recognized that likelihood will sometimes fail to be statistically consistent. The theoretical difference between these positions assumes a practical relevance when we focus on cases in which likelihood inference is inconsistent.

Neyman considers such cases and comes down in favor of the consistency requirement:

> . . . whenever the method of maximum likelihood yields estimates which are both consistent and efficient, this circumstance (but not the principle) may be considered an inducement to use the maximum likelihood estimates. On the other hand, if and when the maximum likelihood estimates are either insufficient or are outside "the pale of decent usage" by being inconsistent, the suggestion to use them, merely because they "maximize the measure of rational belief when we are reasoning from the sample to the population", does not seem convincing. (Neyman 1952, p. 189)

On the other hand, those who believe that it is a first principle that likelihood measures the degree to which hypotheses are supported by the evidence will hold that a likelihood inference may be legitimate even when it is statistically inconsistent.

I rather doubt that this dispute can be resolved by the proof of a theorem. The question before us is not so much technical, as interpretive. In the next section, I address it by constructing an example in which a likelihood rule of inference fails to be statistically consistent. The choice is between using an inconsistent rule of inference or declining to make an inference at all with the information provided. I find it intuitively plausible that the data discussed do offer evidence that differentiates between the hypotheses considered. I will suggest that it is extremely implausible to conclude that the data are devoid of evidential meaning. I therefore reach the conclusion that not all inconsistent likelihood inferences are "outside the pale of decent usage".

This line of argument will not convince all parties to the disputes that have surrounded statistical inference, nor is it intended to do so. Those who deny that data ever provide evidence—that statistics is about deciding what to do, not with inferring what the data say—will not be moved. And even for those who are willing to talk of evidence and inference, my interpretation of the notion of likelihood inference will not be entirely uncontroversial, although it is by no means novel. I use "likelihood inference" as Edwards (1972) and others have done. Likelihood is a device for evaluating how well supported competing hypotheses are in the light of available evidence. It does not tell one to believe that the most likely hypothesis is true and that the other competing hypotheses are false. Likelihood does not provide a rule of acceptance. Likelihood inference culminates in a judgment as to which hypothesis is best supported, not in a judgment as to which hypothesis is true.

Although no prior probabilities will be given for the competing hypotheses considered in the example that follows, Bayesians should not imagine that my examples are objectionable on that score. My argument will not be affected by imagining that the competing hypotheses have identical priors. In this case, Bayes' theorem can be used and posterior probabilities calculated therefrom. This does not affect my reading of the example, since hypotheses with identical priors have different posterior probabilities, relative to given data, precisely when they differ in likelihood. Inconsistent Bayesian inference, for me, can be legitimate for the same reason that inconsistent likelihood inference can.

2. The Example: A Known Probability of Nonconvergence. Suppose a machine producing coins must be in one or another of two states (S_1 and S_2) at any given time and that the state of the machine determines the probability the resulting coin has of landing heads if it is tossed. The machine is so constructed that state S_1 occurs with probability 0.9 and S_2 occurs with probability 0.1.

The hypotheses we wish to test agree that the machine states influence a coin's bias, but disagree on the direction of the influence. H_1 asserts that if the machine is in state S_1, then it produces a coin whose probability of landing heads is 0.8, whereas the machine's being in S_2 gives the resulting coin a probability of landing heads of 0.2. So H_1 says that S_1 biases coins toward heads and S_2 toward tails. The other hypothesis, H_2, makes just the opposite assertion. It says that S_1 gives the coin a probability of landing heads of 0.2 and that S_2 gives the coin a probability of landing heads of 0.8. H_1 thereby says that S_1 biases coins towards tails and S_2 towards heads.

It is easy to design a maximum likelihood rule of inference for this problem that is statistically consistent. One should toss numerous coins drawn at random from the machine. The expectation is that those coins will approximate a 9:1 mixture of coins produced under the two machine states. H_1 implies that the expected frequency of heads in this series of tosses is $(0.9)(0.8) + (0.1)(0.2) = 0.74$; H_2 says that the expected frequency of heads is $(0.9)(0.2) + (0.1)(0.8) = 0.26$. The rule of inference for evaluating what the data say is as follows:

> If the observed frequency of heads is greater than 50 percent, infer that H_1 is better supported; if the observed frequency of heads is less than 50 percent, infer that H_2 is better supported.

This rule is a "likelihood rule" because it judges a hypothesis to be better supported by data precisely when the hypothesis confers a higher probability on that data.

The rule just described is statistically consistent. The law of large numbers implies that if H_1 is true, the sequence of tosses will converge on a frequency of heads of 0.74 and that if H_2 is true, the sequence will converge on 0.26. This satisfies the definition of consistency given earlier: the probability approaches unity that the false hypothesis will be "rejected"—that is, judged less well supported—as the data are made large without limit. Here, then, is a case of likelihood inference with consistency.

However, the problem I want to consider arises when the "ideal" experiment just described cannot be performed. Suppose I hand you a single coin produced by the machine and make it impossible for you to obtain any others. Your task is to repeatedly toss this single coin and reach a judgment about which hypothesis is better supported by the resulting evidence. You are to perform this experiment without knowing what state the machine was in when it produced the coin, knowing only the probabilities of the two machine states.

A likelihood inference, in this case, would take the following form. As before, we compute the expected frequency of heads under the two hypotheses. H_1 says that the expected frequency is 0.74; H_2 says the expected frequency is 0.26. We adopt the likelihood inference rule stated above. However, because we are now tossing a single coin of unknown bias, this rule of inference is inconsistent. The probability of rejecting the false hypothesis is asymptotically 0.9; but consistency requires that this probability of convergence must be unity.

To see why consistency fails in this case, imagine that, unbeknownst to us, the machine was in state S_2 when the coin was produced.[1] In this circumstance, if H_1 is true, then as the number of tosses is made large without limit, the frequency of heads will get arbitrarily close to 20 percent; and if H_2 is true, the frequency of heads will approach 80 percent as a limit. That is, if we are working with an S_2 coin, then whichever of the two hypotheses is true, the above rule is virtually certain to say that *the other hypothesis* is the one better supported by the evidence, as the data are increased without limit.

The probability that convergence fails, in this example, is the probability that the machine was in state S_2, namely 0.1. We may generalize this result a little, in terms of the table displayed below. We do not know which conjunctive hypothesis of the form H_iS_j $(i, j = 1, 2)$ is true. Their likelihoods—the probabilities they confer on the coin's landing heads—are the entries in the table. I stipulated above that $w = z$ and $x = y$.

		Machine States	
		S_1	S_2
Hypotheses	H_1	w	x
	H_2	y	z

As the single coin is tossed repeatedly, its frequency of heads will converge on either $w = z$ or on $x = y$. If this were all we knew about the chance set-up, we would be unable to distinguish between H_1 and H_2. However, we know that the machine was probably in state S_1. This means that if the frequency of heads is quite close to $w = z$ after a very long run of tosses, then H_1 is better supported; similarly, if the frequency is quite close to $x = y$, then we should conclude that H_2 is better supported.

It is crucial to my example that the entries in the above two-by-two table are symmetrical. The asymmetry in the problem is introduced by the difference in probability of S_1 and S_2. As long as these two probabilities differ, a likelihood inference can be made, wherein the probability of nonconvergence is just the lesser of the probabilities of the two machine states.

I have viewed this example as a case of legitimate likelihood inference in which consistency fails. But other interpretations should be considered. One reaction is to claim that the likelihood rule of inference is misguided precisely because it fails to be consistent. This strikes me as too extreme. It seems clear that the sample frequency does have evidential meaning; given the available information, likelihood correctly reflects which hy-

[1] Notice that I talk of the single coin's having a probability of 0.9 of being produced while the machine was in state S_1 and a probability of 0.1 of being produced while the machine was in S_2. A referee has pointed out that those who deny that probabilities ever apply to the single case will balk at this. It should be noted, however, that there are several objective and subjective interpretations of probability that do not view this as objectionable.

pothesis is better supported. It would be preferable, of course, if we did not have to proceed in ignorance of the state of the machine or if we could toss a 9:1 mixture of coins. But this hardly shows that the evidence is uninformative when all we have to go on is the result of tossing a single coin, where we know only the probability that the coin was produced under one or the other of the two possible machine states.

A second reaction might be that my characterization of the example is mistaken. The likelihood rule, so this suggestion runs, *is* convergent after all. After all, to see if a rule is convergent, we must formulate some interpretation of what it means to "repeat the experiment". I took this to mean that the single coin before us should be tossed repeatedly. But it might be suggested that a repetition of the experiment should be understood in a quite different way. Why not describe our use of this single coin as forming part of a hypothetical series in which we toss different coins, 90 percent of which were produced while the machine was in state S_1 and 10 percent while the machine was in state S_2? As noted before, *this* repeated experiment permits a consistent likelihood inference to be made.

So in what does a "repetition of the experiment" consist? We have before us a coin whose bias remains unchanged under repeated tossing. We do not have before us a mixture of coins that have different but unknown biases. Moreover, I stipulated that no further coins are available to us. Our problem is to say how the production process biases coins. The data must be drawn from repeated tosses of this single coin.

I toss this one coin some number of times; there are no other coins available to me. I then ask what would happen if the sample were made large without limit. It is as clear as any counterfactual could be that repeating *this* experiment consists in repeatedly tossing *this* coin (whose probability of heads is presumed to remain constant throughout). It does not mean that I toss this one coin some fixed number of times, then return to the machine for a new coin, which I toss that same number of times, and so on.

Whether a method of inference is convergent depends on the model one uses of the process that generates the data. If one knew the state of the machine when the single coin was produced or if one could toss a 9:1 mixture of differently biased coins drawn from the machine, consistency would be assured. However, if one adopts the model of the experiment I have stipulated to be true—that one repeatedly tosses a single coin of constant but unknown bias, where only the probabilities of the machine states are known—then the assurance of convergence disappears.

I see the interpretation I favor of the above example as steering a middle course between two extremes. On the one hand, I reject the idea that tossing the single coin has no evidential meaning; on the other, I reject the idea that this experiment should be redescribed so that the inference rule comes out convergent after all. The former approach errs in denying that tossing a single coin provides any evidence; the latter retains the connection of likelihood and consistency at the price of giving a false picture of the kind of experiment and inference problem one really faces. In this example, likelihood correctly reports which hypothesis is better supported, even though there is no guarantee that the likelihood rule of inference will converge on the truth.

It should be noted that the inference rule under consideration will be statistically inconsistent even if the probability that the coin was produced while the machine was in state S_2 is only $^1/_{10}{}^{10}$. In this case, it is *almost* certain that the rule will converge on the truth. But *almost* is not good enough, as far as the requirement of consistency is concerned. For consistency demands that the probability of rejecting the false hypothesis should be *unity* as the data are made large without limit.

The demand for consistency strikes me as out of place in a science that aims at making judgments about hypotheses in the face of uncertainty. The reason is that the requirement of consistency is a demand for certainty—although of an extremely hypothetical and asymptotic kind. When faced with the single coin where the probability of S_1 is very high and that of S_2 is very low, the demand for consistency implies that no evidence can be gathered by repeated tossing because the coin *might* have been produced while the machine was in state S_2. One might just as well deny that the senses provide evidence about the world because they *might* be impeded by an evil genius. If S_2 is quite improbable, this suffices for repeated tosses of the single coin to provide evidence, whose import is codified in the likelihood rule stated before.

3. Conclusion. When a rule of inference fails to exhibit convergence, it is sometimes said to be "misleading". This *seems* a just assessment, since in such cases, the rule converges on a false hypothesis as the data set is made large without limit; the rule points to a falsehood and thereby seems to mislead us. Though this seems an innocuous evaluation, I suggest that it is sometimes mistaken. In such cases, it may be *the world* that misleads us. An inconsistent likelihood rule in one experiment may perform its duty as faithfully as a consistent rule may do so in another. In the previous example, the likelihood rule does not mislead but quite correctly reports what the data (and only the data) say.

Likelihood tells us which hypotheses are best supported by the evidence. When the evidence is misleading, the best-supported hypothesis will be a false one. A rule of inference that correctly conveys the evidential meaning of observations *ought* to point to a falsehood when the evidence is misleading. When it does so, it correctly captures what the evidence is saying. When the system generating the evidence is structured so that misleading evidence is overwhelmingly probable, it is hardly surprising that likelihood will point to a false hypothesis when the data set is made infinitely large. But this failure to converge in the long run counts against the likelihood rule no more than does its failure to be infallible in the short run.

As noted before, likelihood, does not provide a "rule of acceptance". It does not say that the best-supported hypothesis ought to be accepted as true. Neither does it provide a magical divining rod that will always (in the long run) point toward the truth, regardless of what the evidence is. Likelihood has much more modest pretensions; it is a "rule of evaluation", simply indicating which hypotheses are best supported by the data. The idea that likelihood codifies "evidential meaning"[2] brings out the analogy between it and a translator of languages, who aims to convey the linguistic meaning of utterances. Likelihood is no more misleading when it fails to be consistent than is an interpreter who correctly reports what an utterance means when *the utterance* is misleading.

Fisher once suggested that likelihood is a primitive postulate; he also suggested that likelihood without consistency is outside the pale of decent usage. Although my arguments do not establish that the first idea is true, I think they do show that the second is false. Perhaps there is some more ultimate set of concepts in terms of which likelihood can be justified. My present point is to cast doubt on the idea that statistical consistency is a criterion that likelihood inference must satisfy. In certain ideal circumstances, the ones enshrined in standard discussions of *MLE,* likelihood inference will be consistent. But this is not a property of all likelihood inference. I have tried to describe an example in which the cost of insisting on consistency is very high. If the idea of likelihood without consistency is rejected, one must deny that certain observations have an evidential bearing on the hypotheses at hand. The intuitive implausibility of this denial is meant to justify the possibility of likelihood without consistency.

Statisticians have almost exclusively focused on the design of optimal experiments. In my first example, one might ask why one should toss the single coin rather than a 9:1 mixture of different coins. The latter is the better experiment, so why talk about the former? The reason is that the world does not always permit one to carry out statistically optimal experiments. I grant that the example discussed here does not touch the idea that consistency is desirable if you can get it (though this more modest

[2] I take this expression, though perhaps not in just the sense in which it was there employed, from Birnbaum (1969).

position raises problems of its own). But situations in which one must perform a nonoptimal experiment or none at all do raise doubts about the claim that consistency is not just a nice property, but a necessary one.

REFERENCES

Birnbaum, A. (1969), "Concepts of Statistical Evidence", in S. Morgenbesser, P. Suppes, and M. White (eds.), *Philosophy, Science, and Method.* New York: St. Martin's, pp. 112–143.

Edwards, A. (1972), *Likelihood.* Cambridge: Cambridge University Press.

Fisher, R. (1938), "Comment on H. Jeffrey's 'Maximum Likelihood, Inverse Probability, and the Method of Moments'", *Annals of Eugenics 8*: 146–151.

———. (1950), *Statistical Methods for Research Workers,* 11th edition. London: Oliver and Boyd.

Hacking, I. (1965), *The Logic of Statistical Inference.* Cambridge: Cambridge University Press.

Kendall, M., and Stuart, A. (1973), *The Advanced Theory of Statistics: Volume 2,* 3rd edition. New York: Haffner Publishing Co.

Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability,* 2nd edition. Washington: Graduate School, U.S. Department of Agriculture.

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate", *Annals of Mathematical Statistics 20*: 595–601.