DISCUSSION

L. J. SAVAGE: Without any intent to speak with exaggeration or rhetorically, it seems to me that this is really a historic occasion. This paper is a landmark in statistics because it seems to me improbable that many people will be able to read this paper or to have heard it tonight without coming away with considerable respect for the likelihood principle.

I, myself, like other Bayesian statisticians, have been convinced of the truth of the likelihood principle for a long time. Its consequences for statistics are very great. A person who after an experiment like those discussed by Birnbaum proposes to use an analysis which is not in conformity with the principle, it seems to me, will have to think quite hard of his excuses for doing so. Nothing is cut and dry. I know some people are not convinced. I know also that some people think, with considerable justification, that there are situations in which the statistical model is not sufficiently well formulated for one to talk of the likelihood function, and such demurrers as that. But not to take the principle seriously no longer seems possible.

The particular emphasis of this paper is to approach the likelihood principle from a different avenue than the Bayesian one. Though I have been interested in personal probability for a long time, I think that I, myself, came to take personalistic statistics, or Bayesian statistics as it is sometimes called, seriously only through recognition of the likelihood principle; and, it took me a year or two to make the transition.

Mr. Birnbaum suggests that maybe not everyone will make the transition. I can’t know what everyone will do, but I suspect that once the likelihood principle is widely recognized, people will not long stop at that halfway house but will go forward and accept the implications of personalistic probability for statistics. In many contexts the two are tantamount to each other. In certain contexts stressed by Mr. Birnbaum they aren’t, because as he says, the problem that he attacked is less comprehensive. In such contexts, the likelihood principle tells us only when two data are equivalent and that’s not enough to really make an inference or a decision. I think that when we come to grips with inferences and decisions, we will find a Bayesian probability sneaking in, as I think I have shown in arguments which are shortly to be published, and which I think are already implicit in this paper.

One may wonder whether there is any difference at all. Does the Bayesian theory go beyond the likelihood principle really? I would be tempted in some way to answer “No,” but I’ve come to think that it really does because the Bayesian principle seems to help systematize the problem of design, the problem of robustness, such as was talked about by Box and Tiao this afternoon, and the problem of non-parametric statistics and decision theories generally, in ways that the likelihood principle alone does not.

An extensive part of this paper is concerned with standard, conventional tests and intervals and the like, as possible interpretations of the likelihood principle. It’s certainly understandable that one would look to things like that when he first encounters the likelihood principle, and other people have, I think, done this sporadically, though not so systematically. My own opinion,
for what it is worth, is that interest in these conventional modes of interpreta-
tion will be a transient phenomena only and that most people will learn to
interpret the likelihood principle in connection with personal probability
measures.

I can't stop without saying once more that this paper is really momentous in
the history of statistics. It would be hard to point to even a handful of com-
parable events.

GEORGE BARNARD:* I'd like to add my greetings, and also congratulations,
to Allan Birnbaum on a fine paper and I would like to welcome him and any
other members who find his arguments convincing to the "likelihood brother-
hood."

Having stated my agreement in principle with what Dr. Birnbaum has to
say, I'd like to first of all qualify it and then perhaps to indicate some more
personal disagreement.

The qualification concerns the domain of applicability of the principle of
likelihood. To my mind, this applies to those situations, and essentially only
to those situations, which are describable in the terms which Birnbaum uses—
that is, in terms of the sample space S, and the parameter space Ω and a prob-
ability function f of x and θ defined for x in S and θ in Ω. If these elements con-
stitute the whole of the data of a problem, then it seems to me the likelihood
principle is valid. But there are many problems of statistical inference in which
we have less than this specified, and there are many other problems in which we
have more than this specified. In particular, the simple tests of significance
arise, it seems to me, in situations where we do not have a parameter space of
hypotheses; we have only a single hypothesis essentially, and the sample space
then is the only space of variables present in the problem. The fact that the
likelihood principle is inconsistent with significance test procedures in no way,
to my mind, implies that significance tests should be thrown overboard; only
that the domain of applicability of these two ideas should be carefully
distinguished.

We also, on the other hand, have situations where more is given than simply
the sample space and the parameter space. We may have properties of in-
variance, and such things, which enable us to make far wider, firmer assertions
of a different type; for example, assertions that produce a probability when
these extra elements are present. And then, of course, there are the decision
situations where we have loss functions and other elements given in the problem
which may change the character of the answers we give.

That is the qualification. Now, the quarrel that I have with the paper relates
to Allan Birnbaum's statement that Fisher and I have advocated the likelihood
principle on the grounds principally that it's self-evident. This I must say is
rather disappointing to me, and perhaps serves to give me a lesson which per-
haps Allan Birnbaum may also draw.

In 1949 I wrote a paper entitled "Statistical Inference" which was clearly so
long that not even Allan Birnbaum managed to read it. It was about the same
length as his own paper! What inference is to be drawn from that I hesitate

* Mr. Barnard was not present but provided a recording of this contribution.
DISCUSSION 309
to say. But I did try in that paper to explain why a principle similar to the likelihood principle, at any rate, should be supported and I don’t think that I effectively assumed the whole principle in the beginning.

There was also a review which I wrote of Wald’s book on sequential analysis in which I think I gave some reasons, for preferring analysis based on the likelihood principle to some of the analyses then current in the literature.

However, these points are rather personal ones and in the main I’m certainly in the warmest agreement with Birnbaum. I very much like the way in which he has attempted to make the likelihood principle clearer and easier to understand for those of us who’ve been brought up in the Neyman-Pearson attitude, using the idea of canonical forms for experiments, which I think represents a major contribution to the literature.

JEROME CORNFIELD: I should like to add my congratulations to those of the previous discussants and to say to Allan Birnbaum that I haven’t quite recovered from the shock of seeing that two principles I had thought reasonable and one which I had thought doubtful imply each other. It is clear that I must either believe all three or disbelieve at least one of the two reasonable ones. What is not clear is on what basis this choice should be made.

One basis for this choice is provided by consideration of a consequence of the likelihood principle—the irrelevance of the stopping rule. This is a consequence that some of us do not find an easy one to accept. In the hope of clarifying the issue, I should like to repeat a question that Peter Armitage asked Cedric Smith, in a discussion of his paper in the Journal of Royal Statistical Society, Series B, 23 (1961).

As Armitage formulated the problem—a Bayesian (to whom of course the stopping rule is also irrelevant) is collecting observations sequentially from a normal population with unknown mean, but with prior distribution uniform from plus to minus infinity. After collecting each observation, he computes the posterior probability that the mean is less than zero and plans to stop when this probability turns out to be some suitably small quantity, say 0.1.

To appraise this rule, one can ask for the probability of stopping and coming to the conclusion that the mean, $\mu$, is positive when it in fact is not. Armitage quotes the law of the iterated logarithm to show that this probability is unity when $\mu = 0$. He then goes on to say that since this probability is on some sense a smooth function of $\mu$, the probability of concluding that $\mu > 0$ when it is in fact some small negative quantity is quite high. Thus, if the stopping rule is regarded as irrelevant, the probability of reaching an erroneous conclusion is large for certain values of $\mu$. This does not seem entirely desirable and suggests that the stopping rule may be more relevant than the likelihood principle says it is. I should like very much to hear Dr. Birnbaum’s comments on this.

IRWIN BROSS: We have heard from the mutual admiration society of Bayesians tonight, but what I have to say is addressed to the experimental statisticians, scientists, and others who might hope to learn something useful about statistical inference from these proceedings. My advice is take it with a grain of salt.
For example, the author claims that his principal result has "immediate radical consequences for everyday practice." However, the derivation has a gross blunder in elementary logic which immediately invalidates the result. The author recommends that "reports of experimental results in scientific journals should in principle be descriptions of likelihood functions, when adequate mathematical-statistical models can be assumed, rather than reports of significance levels or interval estimates."

The gimmick here is the innocent looking phrase "when adequate mathematical-statistical models can be assumed." What does this mean? The author spells this out on p. 6 when he says, "We deliberately delimit and idealize the present discussion by considering only those models whose adequacy is postulated and is not in question." Now there's an enormous difference between postulating a model, which is something that mathematicians do all the time, and postulating the adequacy of a model, which is something mathematicians don't do. Adequacy here refers to a correspondence between a theoretical model and experimental fact, and we are now in the domain of the empirical sciences.

Modern science has an unequivocal directive on this subject. A theory must fit the facts. Any statement obtained by violating this directive has no standing as a scientific statement. Postulating adequacy turns this directive inside out. It says the facts must fit this particular model. Consequently, the scientific value of this recommendation is dubious.

I might add that in any actual scientific study, because of the rule that a theory must fit the fact, the adequacy of a model is always in question, and often the main point for a study is to test the adequacy of a model. From the standpoint of formal logic this means that the class of studies covered by the recommendation which I have read to you is the null or empty class. Consequently, the author's statements are technically correct in the same sense that all square circles are triangular.

To recapitulate, the recommendation has a gimmick that has the logical role of a third postulate. Indeed the other two postulates, clearly fail unless this third postulate is introduced. This third postulate is illegitimate in logic, in science and in common sense as well. So there is no basis for this recommendation.

Without trying to go into detail, I will add that if this recommendation is examined from a practical standpoint, it is very bad advice. It would probably be very little short of disastrous to a scientist who followed it. He probably could not publish his paper if he presented his results in the form of an unintelligible likelihood function. In any case, his colleagues would not know what he was talking about.

Finally, I would like to point out that the basic themes of this paper were well-known to, Fisher, Neyman, Egon Pearson and others, well back in the 1920's. But these men realized, as the author doesn't, that the concepts cannot be used directly for scientific reporting. So, they went on to develop confidence intervals in the 1930's, and these proved to be very useful. The author here proposes to push the clock back 45 years, but at least this puts him ahead of the Bayesians, who would like to turn the clock back 150 years.
Now, I don’t want to imply that current statistical procedures cannot be improved. In many experimental situations the usual models are unrealistic and we do need more realistic foundations for statistical inference.

A number of statisticians, Donald Mainland just to name one, have been working toward developing more realistic models or interpretations of statistical methods. But, this work is almost unpublishable in JASA or other statistical journals.

I particularly don’t want to leave the impression that my remarks are directed against Allan Birnbaum, who is trying hard to be realistic, or against his paper which—despite what I have been saying—is actually head and shoulders above the papers on statistical inference that I’ve read lately.

What I’m really doing here is calling attention to a situation that has developed in recent years where the mathematical statisticians, or some of them, have burned their bridges to reality. They have gone out and played very interesting language games. Now, this is an entertaining thing to do. If you make silly statements, you don’t hurt anybody. If you can trip up on elementary logical points, nobody will notice it. You can juggle your symbols around in a very impressive demonstration of your technical ability at juggling. In short, you can have a great deal of fun.

I don’t want to interfere with the fun that the theoretic statisticians are having. The only thing that concerns me is that perhaps some scientists might take this seriously. Then he could land in serious trouble because he’s not working in an abstract area; he has to deal with experimental reality. If he reports his data in an unintelligible way and doesn’t get his paper published, it hurts.

I don’t want to interfere, as I say, with a game which is quite harmless as long as you do not interpret the words that are used here in their usual sense. The main thing that you have to remember in reading these papers, is that words like “observation,” “experiment,” “design of an experiment”—all these words which have familiar meanings—have completely different meanings to a theoretician. The statements that he makes may be sound for his meanings, but very bad advice if they’re interpreted in terms of ordinary scientific meaning of the words.

So, I advise you to use a certain amount of linguistic care in reading these papers. But go ahead and read them and have a good time.

GEORGE E. P. BOX: I found the remarks of the last speaker very stimulating because, although I pride myself on being a practical person, I have found the exact opposite of what was just said to be true. I believe, for instance, that it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes’ theorem.

The fallacy which it is claimed exists in Birnbaum’s paper is that he assumes the model to be correct. This objection could be made to any approach with precisely equal force. Of course, it is necessary in practice to check the validity of the model, but in the development of the theory it is perfectly sensible to take a particular piece and idealize somewhat in order to study it. This would be
true of any approach. A particular example of the practical importance of the likelihood function is to be found in the problem of nonlinear estimation. Nonlinear estimation has, for some years, been widely used, for example, in the chemical industry. Here we first of all check the adequacy of our model and, having done that, we plot the likelihood function. Experimenters who have, for some time, made considerable advances by use of this technique would be surprised to hear that it was in any way esoteric or impractical.

I. J. GOOD:* My purpose is to argue that this stimulating paper is primarily a contribution to the sociology of statistics rather than to its logic.

It will be adequate for present purposes to confine our attention to discrete probabilities. The principle of sufficiency, (S), can be made more explicit by using the definition of sufficiency. We then get the formulation: (S). Let x be the result of observations and let $t = f(x)$ be a statistic, where x and t can be vectors. If for all $\theta$, $P(x|t&\theta) = P(x|t)$, then t tells us just as much about $\theta$ as does x itself.

Stated explicitly in this way the principle of sufficiency is no longer immediately compelling. It looked more convincing when it was expressed with the help of the technical term "sufficient."

Let $(B')$ be the modern Bayesian principle, that it is legitimate to use the axioms of probability even when this involves the use of probabilities of hypotheses. We have $(B') \rightarrow (L)$, where $(L)$ is the likelihood principle, and the arrow means "implies." Also $(L) \rightarrow (S).

Let $(A)$ be the principle of accepting as true remarks made three times by Fisher. Then $(A) \rightarrow (S)$, since Fisher says that a sufficient estimator includes the whole of the information contained in the observations (Mon. Not. Roy. Astr. Soc., 80 (1920), 758–70; Phil. Trans. Roy. Soc. A, 222 (1921), 309–68; Statistical Methods for Research Workers, Section 3). (It will be clear from what follows that I myself agree with Fisher’s remark, since I am a Bayesian.)

Linguistically, this remark of Fisher’s can be justified by defining $I(G:H) = \log \left( \frac{P(G|H) \cdot P(G)}{P(G|H)} \right)$ as "the amount of information concerning G provided by H." Then it is easy to prove that $I(\theta:x) = I(\theta:t)$, and this justifies Fisher’s remark in terms of this definition of information. (Compare Good, Proceedings Institutes of Electrical Engineers, Part C (3), 103 (1956), 200–4. Note that a non-Bayesian cannot even usually use the notation $I(\theta:x)$.) There are several reasons for identifying the function $I(\theta:x)$ with the evidential meaning of an experiment:

(i) $I$ is the only additive explicatum of information that depends on only probability (see, for example, the above reference);

(ii) If there are only two values for $\theta$, say $\theta_1$ and $\theta_2$, then $I(\theta_1:x) - I(\theta_2:x) = W(\theta_1:x)$, the weight of evidence in favour of $\theta_1$ provided by the observations, that is, the logarithm of the factor by which the initial odds are to be multiplied to get the final odds.

(iii) If there are more than two values of $\theta$, then the weight of evidence in

* A recording of this contribution was heard at the meeting.
favour of any particular value is determined by all the values of $I(\theta;x)$ combined with the initial distribution of $\theta$;

(iv) $I(\theta;x)$ is, apart from monotonic transformations, the only possible explicatum for the power of the "hypothesis" $\theta$ to explain the observations. (See Good, *Journal of the Royal Statistical Society*, Series B, 22 (1960), 319–31.) By the "hypothesis" $\theta$ is meant here the hypothesis that the parameters take the values $\theta$.

We see then that $(B') \rightarrow (F)$, where $F$ is the remark that Fisher made three times. Summarizing:

$$
(B') \rightarrow (L) \rightarrow (S),
(B') \rightarrow (F) \rightarrow (S),
(A) \rightarrow (F) \rightarrow (S).
$$

Allan Birnbaum proves that $(S) + (C) \rightarrow (L)$. Since modern Bayesians and other members of what George Barnard calls the "likelihood brotherhood" already accept $(L)$, Birnbaum's result can win over only those statisticians who find $(S)$ compelling for some other reason, such as by being compelled by the weight of Fisher's authority. Since I cannot see any other reason for finding $(S)$ compelling, I am forced to the conclusion mentioned earlier that the paper is primarily a contribution to the sociology and not to the logic of statistics. Also valuable is the emphasis on the importance of likelihood and other detailed discussion.

Note that when the likelihood depends on several parameters, or even on only one, then it is in the spirit of the reduction of the data (one of the purposes of statistics, as emphasized by Fisher) to replace the likelihood function by an average value. One way of doing this is by using an initial distribution of the parameters, in other words by a Bayesian method.

Since only five minutes are allowed for this contribution, there is scarcely time to take up further the speaker's challenge to Bayesians. Perhaps it will suffice to have used a modern Bayesian philosophy in order to put the speaker's main result into perspective. It also seems worth mentioning that not all modern Bayesians hold precisely the same views. For the latest views of this particular one (Good) see my talk at the Eighth International Conference of Management Science, Brussels, 1961.

On a point of terminology, the expression "Bayes' postulate" has been used in the past to refer to the assumption of a uniform prior (or initial) distribution, and is to be sharply distinguished from "Bayes' theorem" about which there has been much less dispute.

D. V. LINDLEY:* I welcome this excellent paper of Birnbaum's because it clearly recognizes the existence of a problem of inference (or evidential interpretation) in statistics that is not decision theory. The tendency of some writers to think that decision theory is a satisfactory formulation of all statistical problems is regrettable. In inference one is making statements that could be

---

* In Mr. Lindley's absence his discussion was read by Colin L. Mallows.
used to make several widely different decisions in unforeseeable contexts: decision theory, on the other hand, is concerned with a problem in a specific context.

The defect of arguments based on the likelihood function alone, seems to me to be most clearly seen when dealing with nuisance parameters. If \( \theta \) is the parameter of interest and \( \phi \) a nuisance parameter then the likelihood \( L(\theta, \phi) \) is of limited use (and very cumbersome if, as in typical experiments, the dimensionality of \( \phi \) is high). What is required is a function \( \alpha(\theta) \) measuring the "reasonableness" of \( \theta \) in the light of experimental evidence. I use reasonableness here instead of more suitable words like likelihood or probability to avoid prejudicing the issue. A Bayesian can provide such a function, with a simple operational meaning, by integrating out \( \phi \) from the likelihood with respect to its prior distribution. This \( \alpha(\theta) \), for a Bayesian, can then be used in any decision problem that requires knowledge of \( \theta \) for its solution. Furthermore, it is easy for a Bayesian to combine his \( \alpha_1(\theta) \) from one experiment with \( \alpha_2(\theta) \) from another (with different nuisance parameters) to provide an overall assessment of \( \theta \). This combination is very important because it is at the basis of the scientific method. Geneticists in the Soviet Union and Britain ought to agree over experiments concerning a linkage quantity \( \theta \) despite the fact that the nuisance parameters in one place are different from those in another. I do not see how the methods of this paper could enable these things to be done: and I do not understand the argument at the end of \$8.2.

There is a suggestion in \$9 (end of the second paragraph) that one may have to consider whether to use \( \theta \) or some function thereof. Is this not equivalent to the Bayesian's difficulty of whether to use the P.I.R. for \( \theta \) or some function thereof? I agree with Birnbaum that it is necessary to make this decision, and would argue that in reporting on the results of an experiment one should report the posterior distribution of this function (having decided what it is) when the prior distribution of it has been taken to be uniform. The reason for this is that one would then be reporting on \( \theta \) or \( f(\theta) \), in the light of the experiment alone without using prior information. This involves admitting that ignorance of \( \theta \) is not the same as ignorance of \( f(\theta) \). But this is reasonable. I have met binomial experiments in which I was fairly ignorant about \( p \), but certainly not about \( p^{10^6} \); I'd be willing to take large bets that it was less than half.

C. W. CLUNIES-ROSS: We all know that active editing (by suppression of data) can produce results which are misleading when taken at face value. All Cornfield's example on the one-decision rule shows is that we can produce the same results without actively lying.

JOHN W. PRATT: A very important contribution of the paper, as has already been emphasized, is a justification of the likelihood principle. Since I think there are some who do not find completely convincing the conditionality principle on which Birnbaum's justification rests, I'd like to give a direct justification of the likelihood principle which I, at least, find convincing. I will only illustrate the argument.

An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate volt-meter, ac-
DISCUSSION 315

accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean. Later he visits the engineer's laboratory, and notices that the volt-meter used reads only as far as 100, so the population appears to be "censored." This necessitates a new analysis, if the statistician is orthodox. However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all. But the next day the engineer telephones and says, "I just discovered my high-range volt-meter was not working the day I did the experiment you analyzed for me." The statistician ascertains that the engineer would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."

I agree with the engineer. If the sample has voltages under 100, it doesn't matter whether the upper limit of the meter is 100, 1000, or 1 million. The sample provides the same information in any case. And this is true whether the end-product of the analysis is an evidential interpretation, a working conclusion, a decision, or an action.

The argument to here is already at variance with orthodox methods, and the point I wish to make is that if this argument is applied universally, it implies the likelihood principle, as I outlined in a review of Lehmann's book, "Testing Statistical Hypotheses," (Journal of the American Statistical Association, 56 (1961), 163).

I should add that this justification of the likelihood principle uses something like conditioning, though in a different way from Birnbaum. My reason for offering an additional justification is the hope that it may help convince people who still need convincing.

I would like to make a comment about relabelling. I'm not sure of its exact role in Birnbaum's work, but I believe there is more to relabelling than meets the eye when the framework is left abstract. It is not merely a matter of interchanging the labels attached to the states of nature. What is really involved is interchanging the distributions attached to the states of nature. An example, over-simplified to bring it into a two-state framework, would be this. If a certain drug has no effect, it helps the same proportion of patients as a placebo, which, let us say, is 25 per cent; if it has an effect, it helps 40 per cent. Relabelling does not mean that the two states are called 2 and 1 rather than 1 and 2 respectively. It seems to me relabelling gives a situation where no treatment effect means 40 per cent are helped and effect means 25 per cent are helped, instead of no effect meaning 25 per cent are helped and effect 40 per cent. This makes no physical sense to me, and accordingly I certainly don't feel com-
pelled to accept equal prior probabilities in Jeffreys' framework. It is not clear to me how strongly Birnbaum uses relabelling, but I do feel that two samples giving the same likelihood on the same parameter space need not logically have the same evidential meaning unless the physical interpretations of the parameters are identical in the two cases.

I have one final comment, on the practical importance of this kind of discussion. It seems to me anyone who has taught a course in statistics of almost any kind would feel that full acceptance of the likelihood principle ought to change the course entirely.

HOWARD LEVENE: I'd like to apologize to Allan that through my own fault I did not see a copy of his paper ahead of time, so that my comments may be a little long. First of all, I'd like to add my small voice to the attack on Irwin Bross. Dr. Box took care of industry. I'll take care of biology.

In cases where you're interested in the location parameter it may be unreasonable to suppose that you have normality and that the likelihood function tells you everything, but there are so many much more complicated problems in biology where you have to set up a model and then this model dictates what the statistics are. In such cases it seems to me you would be perfectly justified in reporting the likelihood function, and that this likelihood function is interpreted as being part of the general structure and framework in which you have made the original biological problem into a mathematical problem, and if it is said that you are successful in that, you will also be giving useful information about the biological problem.

The second point I'd like to make on the paper itself has to do with the principle of conditionality which sounds exceedingly reasonable as in the example of two instruments of different accuracy and that your analysis should be based on the instrument that was actually used. But, I wonder whether there aren't some conditions being imposed on the application of this conditionality. For example, suppose you were dealing with a randomized plot in agriculture, and you are aware that if you used a systematic design you would be introducing bias to the explanation of varying treatment effects because of differences in fertility of different parts, and accordingly, you'd use a randomized design where you decide by your table of random numbers which plot gets which material.

Here it would seem to me that one could argue against the principle of conditionality, which sounds exceedingly reasonable as in the example of two instruments of different accuracy and that your analysis should be based on the instrument that was actually used. But, I wonder whether there aren't some conditions being imposed on the application of this conditionality. For example, suppose you were dealing with a randomized plot in agriculture, and you are aware that if you used a systematic design you would be introducing bias to the explanation of varying treatment effects because of differences in fertility of different parts, and accordingly, you'd use a randomized design where you decide by your table of random numbers which plot gets which material.

Here it would seem to me that one could argue against the principle of conditionality, which sounds exceedingly reasonable as in the example of two instruments of different accuracy and that your analysis should be based on the instrument that was actually used. But, I wonder whether there aren't some conditions being imposed on the application of this conditionality. For example, suppose you were dealing with a randomized plot in agriculture, and you are aware that if you used a systematic design you would be introducing bias to the explanation of varying treatment effects because of differences in fertility of different parts, and accordingly, you'd use a randomized design where you decide by your table of random numbers which plot gets which material.
THOMAS GOLDMAN: I'm pleased to follow the preceding speaker, particularly since I have something to say on the same topic. I agree with my esteemed teacher, Professor Savage, on the epoch-making character of this paper. I don't feel really qualified to say that I agree or disagree; I just want to sharpen a few issues, I hope, particularly with regard to the principle of conditionality.

Some of the speakers seem to feel that Professor Birnbaum has said that the principle of sufficiency implies and is implied by the principle of likelihood. I am quite sure that this is not the case. Consequently, the principle of conditionality, which if taken with the principle of sufficiency implies, as he says, the principle of likelihood, becomes of crucial importance.

Professor Birnbaum in his paper seems to recognize that this principle of conditionality is something of a stumbling block, and my impression was that he wasn't quite clear as to why anybody should object to it. I think that if you look at it, you can see that there is built right into the principle of conditionality a subsidiary assumption whose presence Professor Birnbaum certainly recognizes explicitly but doesn't focus his attention on. It is one which comes up quite frequently; I shall refer to it as the Probability Mixture Assumption. It's of great importance in such areas as decision theory, game theory, and utility theory, but it is not generally recognized as having the importance I think it does have.

I'd like to reformulate it. The Probability Mixture Assumption can be stated as follows: Given a set $A$ with elements $a$, take a probability mixture of two elements—call them $a_i$ and $a_j$, $i$ not equal to $j$—with probabilities of $p$ and $(1 - p)$; then this probability mixture is itself an element of the set $A$, having all the properties of other elements of $A$. More generally, it is equivalent to the proposition that, if there is a set $S$ with subsets $A, B, C, \text{etc.}$ and elements $x$, and we take arbitrary elements of $S, x_i$ and $x_j$, each of which belongs to the intersection of the same subsets, then a probability mixture, $px_i + (1 - p)x_j$, will be an element of $S$ belonging to the intersection of the same subsets.

Now this is a pretty strong assumption in many instances. It is as though we said, "We know that the difference of two integers is an integer, so we may assume that the difference of two positive integers must also be a positive integer." I think you may agree with me that it is perhaps this Probability Mixture Assumption that gives rise to the problems regarding the principle of conditionality. For example, it implies that if you have a measure $m$ on $A$, so that $m_i$ is the measure on $a_i$, and $m_j$ is the measure on $a_j$, then the measure on the probability mixture is $p$ times $m_i$ plus $(1 - p)$ times $m_j$, but this may not be acceptable to the practicing statistician.

To give an example from utility theory, which is very closely related to decision theory, in the very excellent little textbook by Chernoff and Moses called "Elementary Decision Theory," they attempt in the appendix to give an axiomatic construction of utility theory. You will find that the Probability Mixture Assumption is implicit in their construction. I emphasize the word "implicit" because we all know the danger of implicit assumptions. And in fact, in my opinion, any reasonable person, if confronted with the assumption in
explicit form in that context, would reject the Probability Mixture Assumption, primarily on psychological grounds.

In the present context, the difficulties are not so much psychological as measure-theoretical. Is the probability measure on the ancillary statistic (denoted as h by Professor Birnbaum, equivalent to p in my examples) homogeneous in measure with the probability measure on the experimental outcome x—a, that is to say, of the same evidential quality for purposes of inference?

In conclusion, I would like to urge Professor Birnbaum, before he goes on to erect an elaborate superstructure on these undoubtedly very impressive “Foundations,” to analyze in more detail the implications of his assumptions. Research in this area could be of great value to all of us.

A. P. DEMPSTER: My comments stem from the uneasiness which I habitually feel when confronted with strong, apparently binding consequences of a few simple and plausible postulates. I tend to think that the world really cannot be so uncomplicated.

In this instance I am asked to accept the likelihood principle (L) because it follows from the conditionality principle (C) together with the sufficiency principle (S). But are there not other assumptions and principles which, although stated less explicitly, play important roles in the argument leading to (L)? I would like to single out such a principle, to be called here the uniqueness principle (U). This may be stated as follows.

The uniqueness principle (U): For a given statistician with a given mathematical-statistical model and a given set of observations, there is one and only one evidential meaning.

It would be possible to introduce a stronger version of (U), say (U’), which would require all statisticians to agree on evidential meaning. I do not know whether Professor Birnbaum has (U) or (U’) in mind. It is clear, however, that at least (U) is firmly implicated in the argument of Part I of the paper.

My two chief comments are as follows. Firstly, I would like to suggest that there is no clear positive rationale for accepting (U). Could it not be, for example, that the model and data leave undetermined some ingredient essential to evidential meaning? Or, perhaps, could it not be that the concept of a unique evidential meaning is simply an unrealistic ideal? Secondly, I would like to suggest that leaning too heavily on (U) can only lead to extreme, and therefore unconvincing, standpoints on philosophical statistics. For example, could many statisticians really accept the likelihood principle? (L) implies exclusion from any role in evidential meaning of significance tests, confidence statements, and even so basic a concept as the mean square error of an estimator. To eradicate such concepts from the thought processes of statisticians would require a prodigious brain-washing program. Would such a program really rest on firm philosophical grounds? I, at least, would prefer to relax the uniqueness principle somewhat.

An incidental comment is that Professor Birnbaum appears to relax (U) in Part II of his paper when he allows his “intrinsic” methods to refer to “conventional experimental frames of reference.” The key word is “conventional.” Surely this introduces a logical weakness into the paper. For, if (U) can be
given up so easily in Part II, then how can I find Part I convincing when it relies on \((U)\)? On the other hand, if intrinsic methods are left outside of the canon, then direct non-Bayesian interpretation of likelihoods lacks an intuitive basis, at least to me.

OSCAR KEMPTHORNE:* I was a bit surprised at the emphasis on "experiment," "experimental situation" etc. in Dr. Birnbaum's paper. It is a truism, I believe, that there is never an adequate mathematical statistical model for any actual situation. The "reporting" of "experimental results in journals" of the empirical sciences must of necessity be incomplete not only with regard to the condensation of the data, and the drawing of conclusions from them, but also in the description of the performance of the experiment. And the interpretation of experimental results depends as much on the faith put by the interpreter in the completeness and accuracy of the description of what was done as in what the data indicate.

I found, however, much of interest and stimulation in Dr. Birnbaum's paper. I take the problem to be the purely logical one that the observation \(x\) has a probability \(f(x, \theta)\) and one wishes to characterize the evidence about \(\theta\) supplied by \(x\) and the postulation of the function \(f(x, \theta)\).

I learned much of what I know of statistics by reading Fisher's papers and it is, I believe, implicit in these dating back at least to 1922 (Mathematical Foundations of Theoretical Statistics) that the probability of the observed sample is the basis for any feeling we may have about the parameter's values. It therefore does not strike me as appropriate to cite Fisher 1956 as the reference for Fisher's views on likelihood except that it is evidence that Fisher has not changed his mind over a period of some 30 years or so. It seems obvious to me that the probability (or probability density) with other information such as the origin of the data is all we have, and that the idea of sufficiency arose solely out of considering the probability or the likelihood. I was therefore surprised to see that Dr. Birnbaum finds \(S\) appropriate, and makes it his starting point, when one can arrive at a sufficient statistic only by considering the probability (density) of the sample. It may be that I am quite out of touch with what mathematicians have done with Fisher's concept of sufficiency. The fact that one does not wish to modify a binomial observation by introducing a random variable to smooth out the probability is not to me an argument for considering the sufficient statistic but an argument for not introducing an irrelevancy; or to put things a different way, for not saying I observed \(5+r\) successes in 10 trials when I observed 5 successes.

I found Dr. Birnbaum's paper somewhat obscure as regards the relationships among \(S\), \(C\), and \(L\). He states

(1) "\(S\) is implied mathematically by \(C\)"
(2) Lemma 1: \(L\) implies \(S\)
(3) Lemma 2: \(L\) implies and is implied by \(S\) and \(C\).

In particular, because \(C\) implies \(S\), lemma 2 should read in part, it seems, \(C\) implies \(L\). From the point of view of presentation, one can use \(L \rightarrow C\) and \(C \rightarrow S\)

---

* Mr. Kempthorne was unable to attend. His written comments were communicated to the author and the editor after the meeting.
to get $L \rightarrow S$, and $L \rightarrow C$ and $S$. We do not however appear to know from Dr. Birnbaum's paper whether $S \rightarrow L$ and $S \rightarrow C$, either of which would be adequate to show equivalence of $S$, $C$, and $L$. Is it not the case that Dr. Birnbaum's three points stated above amount to "$L$ and $C$ are equivalent and imply $S$"? In view of Dr. Birnbaum's feeling of the appropriateness of $S$, it would be informative to have more complete information on what $S$ implies. If my summarization is correct, the appropriateness of $S$ alone leads nowhere.

The significance of Lemma 2, that $L$ implies and is implied by $S$ and $C$, depends on what force one is prepared to grant $S$ and $C$ on the one hand or $L$ on the other. I do not find either side to be completely forcing. To consider a simple "experimental" situation, a (possibly naive) interpretation of $C$ indicates that one does not need to know that the data comparing two treatments have arisen from a paired design or a completely randomized design, but such an interpretation is completely repugnant to me, as I presume it would be to R. A. Fisher, whose Design of Experiments book has been dubbed by so many (and I believe correctly) a classic. But perhaps I am stupidly misinterpreting what Dr. Birnbaum says. As regards the force of the likelihood principle, I experience some difficulty in accepting the view that the evidential interpretation of the occurrence of one failure in a sample with prefixed size of 100 is the same as that of the occurrence of the first failure on the 100-th member of a sample of indefinitely large size. The fact that standard error type arguments give different answers in the two cases appeals to me. I wonder if my difficulty arises from the fact that the distribution in question is not continuous. What is the situation under the following circumstances? I am sampling from a normal distribution with mean $\mu$, which I know to be between $-1$ and $+1$, and with unit variance, and I stop sampling when I get an $x$ which exceeds unity. Is the likelihood principle still tenable?

I sympathize with some difficulties Dr. Birnbaum has encountered in quoting or inferring from what Fisher has said in his very long and very productive statistical life. Indeed I would go as far as to say that the search for reconciliation of statements Fisher has made which he undoubtedly feels not to be in conflict but others find conflicting, is the best hope for making sense of the whole situation. I was particularly impressed, for instance, by the following two statements made by Fisher (Sankhya, 23, pp. 3 and 5): "The recognition of the appropriate reference set is an essential step to understanding a test of significance, and, therefore, to setting up an appropriate process of verification" and "The method set out above of generating a sample of the only possible reference set appropriate to the data may be used to illustrate the solution of some other related problems."

Here Fisher seems to be saying that one must have in mind a reference set and that there is only one appropriate such set. I infer from this that "outcomes not actually observed," to use Dr. Birnbaum's phrase, are relevant if they belong to the appropriate reference set. It therefore seems to me that while Dr. Birnbaum seems to make an understanding of Fisher's views an appreciable part of his overall aim, he is a long way from interpreting Fisher's writings.
I have difficulty in reconciling other aspects of Dr. Birnbaum's paper. For example it is stated that $C$ implies $S$, which I interpret to mean that if one accepts the conditionality principle one must accept the sufficiency principle. But Dr. Birnbaum says that $C$ and $S$ together imply $L$. In section 5.8, however, it is noted that $C$ does not lead to a unique answer for binary experiments, whereas I suppose that $L$ would lead to a unique answer. Some clarification would be helpful, in particular as to what specifically in reference [3] is being referred to in section 5.8.

With regard to the latter part of section 7.2, that is, the use of some tail probability versus the use of the likelihood ratio, I am reminded of Fisher's statement (which I do not really understand) that the use of a tail probability [9, p. 66], is "not very defensible save as an approximation." Like most statisticians I have at times been worried about the use of tail areas, i.e. probabilities of outcomes not observed, but at other times have felt that the use of tail areas leads to an easily communicable statement of the form "The probability of getting as large or larger a discrepancy on such and such a hypothesis is 3 per cent." This seems to me a meaningful and relevant statement, particularly in reference to the permutation test of a randomized experiment.

Dr. Birnbaum is not concerned merely with the meaning of fiducial probability, but I would be most interested in his explicit reaction to the following view, which is, I believe, what Fisher has said, perhaps in identical words, and which is directed towards "evidential interpretation."

Suppose we are given that a random variable $x$ is normally distributed around an unknown parameter $\mu$ with unit variance, and that we know nothing about $\mu$. A peculiar position to be sure, but one we might well wish to take in order to say what the observation tells us about $\mu$. Then $\text{Prob}\{x < \mu + k\} = f(k)$ identically in $\mu$. So if we consider the totality of possible $(x, \mu)$ pairs, the frequency with which the inequality $x \geq \mu + k$ is true is $f(k)$. We are entitled to apply the probability based on the complete aggregate to a subset of this aggregate if we have no logical basis for picking out any sub-aggregates for which the probability is different. We must therefore attribute the same probability to the sub-aggregate for which $x$ takes its observed value as to the complete aggregate. In other words the probability that $\mu$ is greater than $x - k$ is $f(k)$, or $\mu$ is normally distributed around $x$ with unit variance. This is exactly what Fisher says, I believe. The statement has been attacked on the basis that if we pick a $\mu$, and get a value for $x$, then $\mu$ is greater than $x$ or less than $x$, and the probability statement would be falsified by actual sampling. Some individuals are surprised that anyone could be so foolish as to make such an obviously false statement. I believe however that these individuals are misunderstanding the statement. It is a logical statement which is unverifiable, because an ingredient in the argument is that we know nothing about $\mu$, and one can certainly not check the statement by cases in which one does know $\mu$. As I understand the situation, the statement that $\mu$ is normally distributed around $x$ with unit
variance is unverifiable by sampling populations with known μ's, i.e. μ's which have no relationship to the x in question, because any attempted verification by random sampling would violate the agreed postulate that we know nothing about μ.

The statement that μ is normally distributed around x with unit variance is a statement of our knowledge about μ after observing x, under the given conditions that x is normally distributed with unit variance around a fixed unknown parameter μ of which we initially know nothing. As such it seems rather appropriate to what Fisher has called the accumulation of natural knowledge. It gives an answer to the question “What is the state of our knowledge on the basis of the postulated conditions with no intrusion of personal feelings about the unknown parameters?”, or what may be loosely termed “What are the facts of the situation?” It is the preeminence of this question which seems to be the real stumbling block for any use of Bayesian arguments in problems of scientific inference except when the sampling of a superpopulation is inherent in the situation. It seems also worth adding that in actuality one may have some more or less vague ideas about the value of μ. One would presumably recover the observations from the fiducial distribution in order to combine them with a prior distribution if such existed, though one might hope that one would not have to retrace one’s steps. Also the fiducial argument does not pretend at all to answer the question, “What should I do on the basis of my observation?” This is a perfectly valid different question for which one must clearly use decision theory, which dates back to the early Neyman-Pearson work, and which requires specification of costs and risks, without which the question posed is empty.

The above reasoning is, I believe, the one and only cornerstone of the whole fiducial argument. If one cannot accept the reasoning there is no point in going to more complicated cases like the “Student” situation or “binary experiments.”

While I may seem to have been somewhat critical of the paper, I believe Dr. Birnbaum has advanced our knowledge appreciably, and I found his paper most stimulating. In particular, he forces us all to re-evaluate what we as statisticians teach and profess.

ALLAN BIRNBAUM: All of the comments have been very interesting and stimulating to me, and I’m very grateful for this opportunity to have heard them.

I’d like to emphasize again, in connection with Dr. Bross’ comments, and those of Professor Neyman (in “Two Breakthroughs in the Theory of Statistical Decision Making,” mimeographed, revised version, p. 30, presented at a December 28 session of these meetings), that I do not “recommend” adoption of (C) and its consequences, nor “disapprove” of methods of interpretation incompatible with (L) or (S). (Bross’ quotation from a paragraph in the middle of Section 1 is incomplete and misleading on this point.) I believe each scientist and interpreter of experimental results bears ultimate responsibility for his own concepts of evidence and his own interpretations of results. My first aim has been to offer formulations of several concepts which seem relevant to a
number of writers, and to examine their relations and implications. My second aim, which is quite distinct, has been to describe how and why, in my small individual capacity as an interpreter of experimental results, it now seems to me that $(S)$, $(C)$, and $(L)$ are appropriate characterizations of statistical evidence in connection with fully parametric models and the delimited purpose of informative inference. No doubt the main value of investigations and discussions like these is their possible help to individual scientific workers in developing their own views and methods.

On a more specific point touched on by Professor Savage, I do not recommend even serious consideration of intrinsic confidence and significance methods as proposed standard working methods; I think the practical value of these is restricted to the conceptual and heuristic spheres, where I hope some others may find them as useful as I have in tracing connections between various concepts and in appreciating that likelihood functions as such are evidentially meaningful and directly interpretable.

The main point about adequacy of models, discussed by Dr. Bross, seems to me very clearly expressed in Professor Kempthorne's initial paragraph and the comments of Professors Levene and Box. Concerning formality of formulations and deductions, discussed by Dr. Bross and Professor Dempster, it seems extremely useful here as elsewhere to formalize concepts when possible, while criticizing proposed formalizations when they seem too broad, incomplete, or otherwise defective. In particular, as to whether one should expect to find an adequate unique characterization of evidential meaning for each formal instance $(E, x)$ of statistical evidence, a point touched on in different ways by Professors Pratt and Dempster, I would like to see more specific grounds for questioning this. The non-uniqueness of intrinsic confidence and significance statements concerns only form, not substance; it is familiar that a physical system can be given different-looking but equivalent descriptions by use of alternative frames of reference. One could replace "the evidential meaning of $(E, x)$" by "the evidential meanings of $(E, x)$" throughout the present paper, as Professor Savage has pointed out in correspondence, without affecting the line of discussion.

As to whether the likelihood position is a halfway house to a Bayesian position and whether one can be expected to go "forward" from it to a Bayesian position, as Savage suggests, I am somewhat skeptical about that but am eager to discuss concrete examples, especially from experimental sciences, in more detail and depth than I have seen done, to see how Bayesian interpretations would appear in comparison with likelihood interpretations combined with unformalized judgments about background information.

Mr. Lindley's question about how nuisance parameters may be dealt with leads to some technical and/or conceptual problems within each general approach to inference; but these problems don't seem crucial to general principles despite their importance and difficulty. My reservations about Bayesian treatments in general also apply here. Questions of robustness of likelihood interpretations, which seem suggested by some of Dr. Bross' comments on adequacy of models, can be formally included among nuisance-parameter problems if we
can at least specify some class of models which we can assume to contain an adequate model, and label the models by respective values of nuisance parameters; then questions of adequacy are subsumed within an enlarged parametric model. Each outcome then determines a likelihood function on an enlarged parameter space; perhaps experience with likelihood interpretations in simple examples will give understanding which will take one surprisingly far in less simple examples. Concrete experience and examples seem to me what is most needed here.

Of course, a large part of the writings of Fisher and of Barnard constitute the predominant part of the present literature on non-Bayesian likelihood methods. As Professors Kempthorne and Barnard pointed out, this paper gives not even a sketch of a comprehensive interpretation of this literature. It is a great pleasure to acknowledge here my considerable debt to Professor Barnard for many very interesting and profitable discussions, as well as to his writings, which include several kinds of formal arguments supporting a likelihood position and many discussions of interpretations of likelihood functions. Some writings of Fisher's, particularly, seem to give much weight to self-evidence (among other considerations, as mentioned in Sections 5.3, 5.4) in supporting the likelihood principle; and this is most impressive to one who has arrived at this principle only by a slow and indirect path.

Professor Pratt's forceful example illustrates another general mode of argument supporting the likelihood principle, which may be compared with the communication channel examples in [3, pp. 418–9.]. It will be interesting to compare a general formulation of this approach with (C).

Mr. Goldman sees in (C) a strong and possibly unacceptable Probability Mixture Assumption, concerned here with existence of mixtures of experiments. But (as emphasized in the examples and Theorem 1 of [3]) the scope of (C) includes a great many familiar models of experiments which we find, by mathematical analysis, to have (generally non-unique) mixture structures; at least in these cases there is no question of having to postulate existence of mixtures. A somewhat restricted deduction of (L) from (C) can be based on such cases alone; but it also seems natural for purposes of general formal discussion to consider simply all experiments (including all mixtures of experiments).

Professor Kempthorne distinguishes between "considering the sufficient statistic," and "not introducing an irrelevancy" such as a randomization variable. But most familiar models contain the equivalent of randomization variables "built in," while other formal models really do contain randomization variables added by an experimenter for whatever reason. I see no concept except (S) which characterizes irrelevancy appropriately for purposes of general formal discussion.

He correctly summarizes "L and C are equivalent and each implies S." And (S) alone does not imply (L), since (S) concerns only different outcomes of any single experiment. (His point (2) misinterprets Lemma 1 as stating "L implies S." The latter is true, but the lemma concerns not the likelihood principle, but the likelihood function in its technical relation to sufficiency. Throughout the paper "(C) and (S)" could be replaced by "(C)"; this was not done because a formal proof that (C) implies (S) did not seem worth adding to the paper.)
On Professor Kempthorne's comment about whether one needs "to know that data comparing two treatments have arisen from a paired design or a completely randomized design," (C) as formulated applied only to fully specified ("parametric") models, for which each outcome determines a fully specified likelihood function. As Professor Kempthorne's writings illustrate abundantly, many models which usefully represent randomization concepts and techniques are not so fully specified. Nevertheless, it seems interesting to try to trace the relations between (C) and randomization in experimental design, and evidently one way of doing so is to supplement a randomization model, at least for purposes of such discussion, so as to obtain a complete parametric model. To illustrate in miniature why (C) seems to me compatible with the role of randomization, let \( D_1 = \Delta + \beta + e \) be the model of the difference between an observation representing treatment 1 and an observation representing treatment 2; the parameter of primary interest is the unknown difference of treatment means, \( \Delta \); \( e \) is a standard normal measurement error; \( \beta \) is the unknown bias due to plot effects when treatments 1, 2 are applied to plots 1, 2, respectively; call this design \( E_1 \). Let \( E_2 \) differ only in that treatments are assigned to plots in the opposite order; then the model for \( E_2 \) is \( D_2 = \Delta - \beta + e \). Let \( E^* \) be the mixture of \( E_1 \) and \( E_2 \) with equal weights; any outcome of \( E^* \) is represented by \( (E_1, d) \), where \( d \) is an observed value of \( D_1 \), or by \( (E_2, d) \), where \( d \) is an observed value of \( D_2 \). Let \( E \) be the censored version of \( E^* \) in which an observed \( d \) is reported without the information "\( E_1 \)" or "\( E_2 \)"; then each outcome \( d \) of \( E \) determines a likelihood function

\[
 f(d, \Delta, |\beta|) = \frac{1}{2} \phi(d - \Delta - \beta) + \frac{1}{2} \phi(d - \Delta + \beta),
\]

where \( \phi \) is the standard normal density. The algebraic sign of the bias \( \beta \) may be regarded as one of two nuisance parameters (the other being \( |\beta| \) which was present in designs \( E_1, E_2, \) and \( E^* \), but which is absent from the randomized design \( E \). In \( E \) there is no ancillary statistic by which conditional frames of reference could be defined; thus no question of compatibility between (C) and randomization seems to arise.

The questions about alternative sampling rules are recurring ones in discussions of (L), and for this reason the example of Section 7.3 was prepared to illustrate the roles of (C) and (L) in detail in a concrete context. Such examples have helped some, including the writer, and I think they can help others, to resolve such questions to their own satisfaction.

Professor Kempthorne's questions about Section 5.8 may be answered by the examples and discussion adjacent to Th. 1 in [3], and by noting that it is only when (C) is applied in the "usual" restricted way that it gives non-unique "answers," while (C) interpreted generally leads to (L).

In reply to Professor Kempthorne's question on the fiducial argument, I suspect that I don't understand the meaning ascribed to fiducial probabilities even in simple examples where I see how they have been computed. For this reason I offered in Section 11 an attempt to interpret the fiducial argument in a way compatible with (L), on which I hope that those interested in fiducial probability will comment.

I take the essential issue in Mr. Cornfield's example to be not the Bayesian
form of defining a sampling rule, but the question (also commented on by Mr. Clunies-Ross) whether likelihood interpretations may tend to be misleading in connection with certain sampling rules or experiments. This issue may be illustrated by an experiment in which \( Y = 1 \) with probability .99 under \( H_1 \) and with probability .9999 under \( H_2 \), \( Y = 0 \) otherwise, and just one observation on \( Y \) is available. \( y = 0 \) would give likelihood ratio .01, "fairly strong evidence for \( H_1 \) against \( H_2 \)"; this occurs rarely under either hypothesis, more rarely under \( H_2 \). \( y = 1 \) would give likelihood ratio 1.01, "evidence supporting \( H_2 \) over \( H_1 \) so slightly as to be practically uninformative"; under each hypothesis this is nearly certain to occur, slightly more so under \( H_2 \). The fact that under \( H_1 \) the probability is .99 of an outcome which will be interpreted as evidence qualitatively supporting \( H_2 \) does not seem unreasonable, when taken with the almost negligible strength ascribed to that evidence, and with the .01 chance under \( H_1 \) of getting evidence quite strongly supporting the true hypothesis. Of course such considerations should whenever possible be made at the stage of experimental design, when one can control probabilities of misleading or uninformative outcomes by determining appropriate sample sizes, etc.

For further clarification of these questions, I look to additional critical study and discussion of old and new arguments, formal and informal, and of concrete examples, especially in detailed empirical research contexts.