THE USE OF THE LIKELIHOOD FUNCTION IN STATISTICAL PRACTICE

GEORGE A. BARNARD
IMPERIAL COLLEGE

1. Introduction

The title I have chosen deliberately echoes that of the paper by L. J. Savage, [1] because it is written with an objective which closely corresponds to Savage's —to encourage practical statisticians to explore the ways in which the study of the likelihood function generated by a set of data can help in its interpretation; At the same time I hope theoretical statisticians will be encouraged to study the theory of likelihood with a view to explaining in detail how the likelihood function can be used, and what its limitations are. It appears to be high time we did this, because for some years now it has been common for geneticists to express themselves in terms of likelihood, and the following quotation indicates that high energy physicists are following suit: "How then can an experimenter present the results of his work in an 'objective' fashion, that is, without introducing his own prior beliefs? One way, *often used by physicists* (my italics, G.B.), is to present $L(x_i^{obs}; \alpha)$ as a function of α for his particular observations $x_i^{obs}; \ldots$" [2].

Whereas in some fields it is still possible to attribute failure to use "orthodox" statistical methods to mere ignorance, such an explanation is untenable in the case of statistically sophisticated areas such as these two. It must be here that the "orthodox" methods have been tried and found wanting.

Lest my allusion to Savage be misinterpreted to mean that I accept the subjective Bayesian position, let me hasten to specify some of the ways in which we differ. Whereas Savage, if I understand him aright, would regard a specification of the likelihood function as *always* providing, at least in principle, the solution to problems of statistical inference, I conceive of likelihood methods as rigorously applicable only to those situations where the distribution of the observations x over the sample space S can be taken as known to belong to a (usually continuously) parametrized family of distributions with probability functions $f(x, \theta)$, the parameter θ ranging over a well-defined parameter space Ω. The primary problem in such a case is often how to express the order of preference among the different values of θ which may be said to be rationally induced

This paper was written while the author was a Visiting Professor at Stanford University, attached to the Computing Group, Stanford Linear Accelerator Center, under the auspices of the United States Atomic Energy Commission.

when the observations x are known. It is this problem which, it seems to me, is answered by giving the likelihood function

$$L(\theta) = f(x, \theta)/\sup_{\theta} f(x, \theta). \tag{1}$$

This problem is different from that of establishing any particular one of the θ's as "true" in some absolute sense, or of rejecting any such particular value as incredible. Being equipped with a likelihood function is like being equipped with a balance, but no weights, and being confronted with several specimens, $A_1, A_2, A_3, \cdots; B_1, B_2, B_3, \cdots; C_1, C_2, C_3, \cdots$ of each of several denominations A, B, C of coins. With such a balance we could establish, for example, that a coin of denomination A is more than twice, but less than three times as heavy as one of denomination B. But we could not express the weight of any of the coins in grams. And the analogy with the balance would be misleading, unless we imagined that the coins have the peculiarity that different denominations react chemically to dissolve into thin air when put together on the same scale pan. It would then have no verifiable meaning to say that the weight of a coin of denomination A is equal to the combined weight of a coin of denomination B put together with a coin of denomination C. Correspondingly, if the likelihood of θ_1 against θ_2 is $\frac{4}{1}$, while that of θ_1 against θ_3 is $\frac{4}{2}$, it will have no meaning to say that the likelihood of θ_1 against "θ_2 or θ_3" is $4/(1 + 2) = \frac{4}{3}$, because the likelihood of "θ_2 or θ_3" is not well defined on the data x.

If each of θ_2 and θ_3 had a definite prior probability, and these prior probabilities stood in the ratio $p:(1 - p)$, then we could interpret "θ_2 or θ_3," in relation to x, as meaning "θ_2, with probability p, or θ_3, with probability $1 - p$." Then the likelihood for "θ_2 or θ_3," given x, would be proportional to $pf(x, \theta_2) + (1 - p)f(x, \theta_3)$. My unwillingness to suppose that such combinations of likelihoods always have meaning is another respect in which my position differs from that of the subjective Bayesians.

But perhaps the major difference lies in my view that there are inference problems not of this comparative form, where it *cannot* be taken as known that the distribution of x belongs to a well-defined family. In relation to an hypothesis H, it may well be appropriate to define a (real-valued) measure of discrepancy $T(x)$ and, given an observed result x_0, to calculate the probability $\Pr\{T(x) \geq T(x_0)|H\} = \alpha(x_0)$. If this probability is small, we shall be faced with the disjunction: either an event of small probability has occurred, or H is false. Thus with sufficiently small α, we shall be disposed to think up some alternative to H which fits the observations better.

When, in such a case, we say that H is rejected on the data at the α level of significance, we are using α as a "measure" of credibility of H, on the data, in the simple but useful sense, that as α ranges from 1 down to 0 our disposition to think of some alternative will increase. On the other hand, of course, if H is rejected by data x on the α level of significance, while another hypothesis H' is rejected by other data y on the α level of significance, it will not at all follow that our disposition to think of an alternative to H' must be the same as our disposition to think of an alternative to H. Only if H' is rejected by the same data, using the same criterion T, on a higher level of significance than H will it be illogical, in the absence of other evidence, to accept H' while rejecting H.

It is an elementary error, of course, to think of Pr $\{T(x) \geq T(x_0)|H\}$, when it is known that $T(x) \geq T(x_0)$, as being the *probability* of H. It is a measure, or ranking, of credibility of a different, cruder, but nonetheless useful kind. Likelihood may, in a loose sense, be thought of as standing intermediate between this crude measure of credibility, and the very precise measure provided by a statement of probability.

I should, perhaps, make clear that I am making no serious claim to originality in this paper. It might well be regarded as a sermon on the text, from an early edition of Fisher's *Statistical Methods:* "The mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences (from sample to population—G.B.), and . . . the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term 'Likelihood' to designate this quantity*; since both the words 'likelihood' and 'probability' are loosely used in common speech to cover both kinds of relationship. . . ."

"*A more specialized application of the likelihood is its use, under the name of 'power function', for comparing the sensitiveness, in some chosen respect, of different possible tests of significance."

I shall try to expound what I take this to mean.

2. The likelihood function as an operating characteristic

My first point is a gloss, or extension, on Fisher's footnote. One of the great services that have been rendered by Neyman and Pearson to mathematical statistics is the emphasis they have placed on the operating characteristic of a statistical procedure. First applied by them in connection with hypothesis testing, or two-way decision procedures, the concept has more recently been generalized to cover multiple decision procedures. If the sample space S is finite (as we shall later argue is really *always* the case) and there are k possible decisions D_i ($i = 1, 2, \cdots, k$), and if C_i is the region in S where we take decision D_i, then the operating characteristic of our decision procedure specifies Pr $\{D_i|\theta\}$ as a function of θ; that is, it gives

$$P_i(\theta) = \sum_{x \in C_i} f(x, \theta) = \sum_{x \in C_i} w(x) L_x(\theta), \tag{2}$$

where $L_x(\theta)$ is the likelihood function, given the observation x, and

$$w(x) = \sup_{\theta} L_x(\theta), \tag{3}$$

the normalizing factor of the likelihood function, may be regarded here as a (nonnegative) weighting factor. In other words, *the i-th component of the operating characteristic $P_i(\theta)$ is a weighted sum of the likelihood functions for the points leading to the i-th decision.*

(Here we take the operating characteristic to consist of the set of functions which gives the probability of taking the i-th decision as a function of the parameter value θ. This is a more direct generalization of the original idea of the power curve than the (single) function which gives the mean value of the loss as a function of θ. The former becomes a special case of the latter, of course, if we allow vector-valued loss functions. But the former, though less general, has the advantage that it retains a precise meaning in situations, often arising in practice, where the exact evaluation of losses is impossible.)

We can go further if we restrict consideration to decision procedures which are admissible. For, under wide conditions, we know that such procedures must be Bayes with respect to some prior on Ω, and hence have the property that two points x, x' in S for which the likelihoods $L_x(\theta)$, $L_{x'}(\theta)$ are the same, must lead to the same decision.

Thus, in such conditions we can say that any component of the operating characteristic of any admissible procedure is obtainable by forming weighted sums of possible likelihood functions.

Now let us introduce a new notion—that of a *nonrestricting* class of possible decisions. The idea behind this is that when the number k of possible decisions is small we often, in practice, feel that we should be able to have a wider choice; and it often happens that when we review the possibilities open to us, we are in fact able to choose from among a wider class of decisions than we thought. For example, we often find the simple two-way decision alternatives in an hypothesis testing problem restrictive—to declare rejection at the 5% level, or nonrejection, say—and we may give ourselves three alternatives: rejection at 5%, rejection at 1%, or nonrejection. Or, again, we may take as a third alternative, in such a case, to continue sampling.

The artificiality of a restricted class of decisions is perhaps most marked in connection with procedures such as those which have been developed to select the k largest from a set of m means. It can easily happen that we find ourselves with clear evidence that $k - 1$ of the means are larger than the rest, but we really have no substantial grounds for selecting the k-th one from the remaining $(m - k + 1)$. Then we may well regret having said we would choose k, and would prefer to be allowed to choose only $k - 1$. Or again, we may find that $k + 1$ of the means are clearly larger than the rest, but there are no appreciable differences between these. Here we would wish to be allowed to choose $k + 1$ instead of k; and this may, on re-examination of the real situation, prove to be allowable in such a case. To represent the practical situation fully, we would need to set up the problem as one of selecting *approximately* k means, with a loss function for departure from k, as well as a loss function for selecting a smaller rather than a larger mean. If we do not do this, it is only because the theory of such a problem could well become unmanageable.

On examination, it appears that the difficulties we experience, in situations such as these, arise from the fact that limitations on the alternatives allowed to us, force us to make the same decision for two distinct observations x, x', even though the information which x provides about θ is different from that which x' provides, in the sense that the value of any minimal sufficient statistic for θ is different at x from what it is at x'. A nonrestrictive class of decisions may then be associated with an admissible decision rule which is such that the class C_i of observations x which lead to decision D_i consists of those observations, and only those observations, for which a minimal sufficient statistic takes a given value. But in such a case, the class C_i will consist of those observations, and only those, for which the likelihood function is a fixed $L_i(\theta)$. And then $P_i(\theta)$, which we have seen is a weighted sum of likelihood functions, will evidently be proportional to this $L_i(\theta)$. Thus, in this sense, a decision function which makes full use of the information provided by the observations will have an operating characteristic whose components are proportional to the set of possible likelihood functions.

It is important to notice that this result is "robust under approximation," in the following sense: we may say that a set of possible decisions is not *seriously* restrictive if it does not force us to make the same decision for two observations x, x' for which the information about θ is *very* different; for this last will mean that the likelihood functions $L_x(\theta)$ and $L_{x'}(\theta)$ will be approximately equal, and thus their weighted sum will be approximately equal to either one of them. Thus, when the set of possible decisions is not seriously restrictive, the i-th component $P_i(\theta)$ of the operating characteristic of any admissible decision rule will be approximately proportional to the likelihood function $L_x(\theta)$ for any x in C_i.

Therefore, one reason for looking at the set of possible likelihood functions which can arise from a given experiment is that it gives us a conspectus of the set of operating characteristics of admissible decision rules. Hence, it may well lead to economy of thought to look at these likelihood functions, especially in the many practical situations where it is not at all easy to obtain a true conspectus of the set of all possible decisions, together with their associated loss functions.

To conclude this section we may note that presumably what Fisher had in mind in writing his footnote was that the choice of a critical region C in S, together with the decision rule to reject the hypothesis tested if x fell in C, and not to reject it if x fell in $S - C$, was tantamount to introducing a new observable y, taking the value 1 if x fell in C, and the value 0 if x fell in $S - C$. The sample space S would then be mapped onto a two-point sample space $\{0, 1\}$. The likelihood function of θ, given $y = 1$, would then be proportional to the power curve associated with C; and indeed, if, as would often be the case, $P(\theta)$ had supremum 1, the likelihood function and the power curve would be identical.

3. Frequency interpretations

One of the red herrings most frequently drawn across discussions on the foundations of statistics is based on the false idea that some measures of uncertainty can be given frequency interpretations, but others cannot. An associated fallacy, against which Charles Stein has spoken out clearly, but few others have, is the idea that such frequency interpretations are unique. It is worthwhile stressing the universality, and the nonuniqueness of frequency interpretations, because this helps to get rid of the idea that the *meaning* of measures of uncertainty is to be found in a particular set of frequency interpretations. The meaning of ideas like probability, significance level, and likelihood, is really to be found only in their use; frequency interpretations are expository only.

Thus the commonest frequency interpretation of a level of significance asks us to imagine a long series of cases, in each of which the hypothesis tested is true. Then if, whenever significance level α is attained, we say that the hypothesis tested is false, the long run frequency of errors is α. More precisely, and a little more generally, if we use the same rule for assertion in a long series of cases, in which the hypothesis tested is sometimes true and sometimes false, the long run frequency of errors cannot exceed α. But such an interpretation of 'significance level α' runs into difficulty when we carry out multiple tests on a single set of data. For this reason, among others, I personally prefer an account which runs as follows.

Imagine that, at any particular time, one's view of the world is represented by the tentative acceptance of a number of hypotheses H_i, each relating to a set of data x_i, the sets and the hypotheses being reasonably independent of each other. If we have a measure of discrepancy (or test criterion) T defined for each set of data, in relation to its associated hypothesis, the level of significance associated with x_i, in regard to H_i, is

$$\alpha_i = \Pr\{T \geq T(x_i)|H_i\}. \tag{4}$$

Now the α_i, if all the H_i were true, would be distributed (nearly) uniformly over the interval [0, 1]. For example, roughly 1% of the α_i should be less than 0.01, and if the actual proportion is greatly in excess of this, we should consider modifying some of the associated H_i. It would also call for some consideration if far fewer than 1% of the α_i were less than 0.01, though attention would in this case first be directed at the formation of the criterion T, or the calculation of its distribution. Hypotheses giving low values to α should thus be regarded much as luxuries by a person of moderate means—not to be avoided entirely, but not to be indulged in to excess.

But neither this nor the more common frequency interpretation of significance level is to be regarded as more than expository. The meaning of "H is rejected at significance level α" is "Either an event of probability α has occurred, or H is false," and our disposition to disbelieve H arises from our disposition to disbelieve in events of small probability.

In the same way, to say that, on data x, θ is k times as likely as θ' *means* that $L(\theta)/L(\theta') = k$, simply; but we can give such a statement a frequency interpretation as follows: imagine a long series of experiments, in each of which we have to decide between two alternatives such as θ and θ'. If, in such a series, we assert the truth of the more likely hypothesis only when the likelihood ratio (such as $L(\theta)/L(\theta')$) exceeds k, then the "odds against error" are bounded below by k. Here we define "odds against error" as the ratio

$$\frac{\text{Long run number of times we rightly decide}}{\text{Long run number of times we wrongly decide}}. \tag{5}$$

(As was first noted by C. A. B. Smith, the truth of this interpretation follows from a simple application of Markov's inequality to the nonnegative random variable L/L', which has mean value 1 on θ'.) We can thus attach an appropriate form of "inductive behavior" to likelihood ratios, if we wish to. But in truth, as with significance levels, our disposition to prefer θ to θ' (assuming k greater than 1) arises from our disposition to prefer that hypothesis which makes more probable what we know to be true. In this case, the disposition to prefer is measured in the rather precise sense, that further independent data having a likelihood ratio less than $1/k$ would need to be forthcoming before our preference between θ and θ' was reversed.

Another form of frequency interpretation for $L(\theta)$ can be given, and is useful, in cases where Ω can be regarded as a separable topological space, in which $L(\theta)$ is always continuous, with a nonnegative measure m defined on open sets. It is then possible to define a sequence $T = \{\theta_r\}$, $r = 1, 2, 3, \cdots$ of points in Ω such that, for any pair of open sets A, B in Ω,

$$\lim_{n\to\infty} \frac{\text{Number of times } T \text{ visits } A, \text{ in first } n \text{ terms}}{\text{Number of times } T \text{ visits } B, \text{ in first } n \text{ terms}} = \frac{m(A)}{m(B)}. \tag{6}$$

If, now, we imagine a long run of cases, in which the true value of θ on the r-th occasion is θ_r, and select from such a long run all those cases in which the likelihood function was $L(\theta)$, then in the subsequence so selected, for any pair of open sets A, B,

$$\lim_{n\to\infty} \frac{\text{Number of times the subsequence visits } A, \text{ in first } n \text{ terms}}{\text{Number of times the subsequence visits } B \text{ in the first } n \text{ terms}} = \left(\int_A L(\theta)\,dm\right)\Big/\left(\int_B L(\theta)\,dm\right) \tag{7}$$

and if, in particular, $\int_\Omega L(\theta)\,dm$ is finite, then $\int_A L(\theta)\,dm/\int_\Omega L(\theta)\,dm$ is the relative frequency with which the subsequence visits A. (These statements are true, with probability 1, in the infinite product space $S \times S \times S \times \cdots$, in which the measure in the r-th component is defined by $f(x, \theta_r)$. They follow from a countable number of applications of the strong law of large numbers.)

This last form of frequency interpretation may be contrasted with that commonly given to a set of confidence intervals $C(x)$, for a parameter θ, with confidence coefficient $1 - \alpha$. Here we imagine the rule $C(x)$ applied to every one of a sequence of cases, in which θ varies in any manner, and we can say that the relative frequency with which $C(x)$ contains the true value of θ is $1 - \alpha$ with probability 1. Now, whereas this frequency interpretation is more general than that just previously discussed, it is open to the objection that it may require us to group together cases where $C(x)$ consists of the whole of Ω, when the true value is certainly in $C(x)$, and cases where $C(x)$ is quite small, and it appears most implausible that θ should be in $C(x)$. The truth of the statement of long run frequency cannot be doubted, but it is hard to justify its relevance to individual cases, when they fall into such discrepant sets. (That cases where this happens are by no means pathological is indicated, for example, by the problem of the ratio of two normal means.) It would seem that lumping such cases together is just as 'arbitrary' as choosing a particular measure m, especially if the value of the ratio

$$\int_A L(\theta)\,dm \Big/ \int_\Omega L(\theta)\,dm \tag{8}$$

does not depend critically on the choice of m.

Perhaps it may be well to spell this out somewhat. By 'arbitrary,' we mean 'dependent on an act of will.' Now the distinguished author of the theory of confidence intervals has emphasized that his notion of 'inductive behavior' involves an act of will. And I am now suggesting that this act of will is one which not all of us would in fact make, since it may involve us in grouping together statements about parameters, some of which are obviously true (since they assert merely $\theta \in \Omega$), while others are obviously doubtful (since they assert '$\theta \in A$,' where A is a comparatively small subset of Ω). The act of will involved in collecting together statements of the form '$\theta \in A$' and statements of the form '$\theta \in \Omega$,' and labeling them all with the same confidence coefficient is not one which all of us will wish to make. In the same way, if we refer a statement of the form '$\theta \in A$,' on the basis of a given sample, to a sequence of cases satisfying the conditions of the measure m above, and so attach to this statement a 'measure of credibility' given by the above formula, we are also making an act of will (in choosing m), but one which, in many cases, could be less objectionable than that involved with the confidence interval.

The objection here leveled against confidence intervals is much weakened, if not removed, by inverting the usual form of statement. If, instead of asserting '$\theta \in A$,' with confidence coefficient $1 - \alpha$, we say "any value of θ in the complement $-A$ of A is rejected at significance level α," those cases where the confidence set covers the whole parameter space lose their absurd appearance; since we are then merely saying that the data are insufficient to enable us to reject any of the possible values of θ. And even if, as many of us feel, it is inappropriate to try to inflate the 'significance' of precise experiments, by adding to the denominator of the frequency a number of experiments which were insufficiently precise, we still have recourse to the simple disjunction, "either θ is not in $-A$, or an event of probability $\leq \alpha$ has occurred."

4. An example from particle physics

It is fundamental to the point of view of the present paper that the theory of statistical inference exists to serve those who are collecting and using empirical data, either to further natural knowledge or to improve the making of decisions; and it is more especially (though not exclusively) with the furtherance of natural knowledge that we are here concerned. It follows that our judgment of the value of any proposed statistical technique will be based on the way in which the technique in question appears to meet the purposes of those engaged in this pursuit. It seems worthwhile, therefore, to consider a specific scientific situation in which these ideas come into play. The broad nature of the problems facing workers in particle physics is sufficiently well known to provide a useful example.

If a beam of polarized Λ particles is observed to decay into a proton and a pion in a bubble chamber, it is possible for each such particle to measure the cosine x of the angle between the track of one of the decay particles and the direction of polarization of the decaying particle. It is then known that observations on distinct particles are independent of each other, and each x follows a distribution with density

$$f(x, \theta) = \tfrac{1}{2}(1 + \theta x), \qquad -1 \leq x \leq +1, \tag{9}$$

where θ is the product of the degree of polarization of the decaying particles, and the parity nonconservation parameter.

A set of n observations $x_1, \cdots, x_i, \cdots, x_n$ then gives a likelihood function $L(\theta)$ proportional to

$$\prod_i (1 + \theta x_i) = 1 + s_1\theta + s_2\theta^2 + \cdots + s_n\theta^n \tag{10}$$

where the s_r are the elementary symmetric functions formed by taking the sum of products of the x_i, r at a time. Since the likelihood is a polynomial in θ with real roots, all of them outside the interval $-1 \leq \theta \leq +1$, it is easy to see that unless all the x_i have the same sign (an event whose probability depends on θ, but which cannot happen with n as small as 10 in more than about 5% of cases), there will be a unique maximum of $L(\theta)$ lying between the largest negative root of $L(\theta)$ and the smallest positive root. It will turn out more convenient to deal with the logarithmic derivative of the likelihood,

$$g(\theta) = \partial(\log L(\theta))/\partial\theta = \sum_i (x_i/(1 + \theta x_i)), \tag{11}$$

which is obviously monotone decreasing, with probability 1, in all cases. We shall argue that for most purposes the best summary that can be provided of the data consists in specifying those values of θ for which $g(\theta)/g'(\hat{\theta})$ takes given values—if one is restricted to giving three numbers, then the roots of $g(\theta) = 0, \pm 2g'(\hat{\theta})$ will probably be most useful, provided that they all lie between -1 and $+1$; while if they do not, instead it will be better to quote the value of $g(1)$, or of $g(-1)$, or both, as the case may be. If more than three numbers can be given, then the roots of $g(\theta)/g'(\hat{\theta}) = 0, \pm 2, \pm 3$, with corresponding provisos, will be best. We shall suggest that ideally, of course, it would be best to specify all values of $g(\theta)$ in the interval $(-1, +1)$ if this were practicable. When the root $\hat{\theta}$ of $g(\theta) = 0$ lies in the interval, it is the value which maximizes the likelihood, and $L(\theta)$ can be recovered as

$$L(\theta) = \exp \int_{\hat{\theta}}^{\theta} g(\theta)\, d\theta. \tag{12}$$

It is easy to see how $L(\theta)$ could be recovered in other cases also.

Now why should we suggest that an approximate specification of $g(\theta)$ is the 'best' way of summarizing the data? Such a judgment must involve an assessment of the purposes for which the data are being collected. In this case the main object is to collect numerical data on the properties of 'elementary' particles, with a view to finding some underlying regularity. The situation is very similar to that obtained in chemistry around the time when Newlands, Mendeleyev, and Lothar Meyer worked towards establishing the periodic classification of the chemical elements, when poor estimates of some atomic weights as well as the fact that some elements had not yet been found, gave rise to difficulties in establishing the classification—indeed a case could be made out for the thesis that the classification was established just as soon as the numerical properties of the elements were established with sufficient precision. With the 'elementary' particles of today it may well be that some underlying regularities are also being obscured by inaccuracies in the data.

The function of the data analysis, then, is to *point towards* some values for the masses, decay cross-sections, and so on, of the various particles, as being more plausible than others. At the same time, the analysis should indicate how far a theoretical value may differ from an experimental value, without having an overwhelming weight of evidence against it. An important feature is, that we should be able in some way to associate a 'weight of evidence' with each deviation of theory from experiment, in such a way that the combined deviations can be assessed. In practice this is often beset with computational difficulties, which may mean, for example, that only first-order perturbation values are available from theory; but should the extent to which a theory fits other facts appear to justify it, further calculations are usually undertaken. Thus there will usually be an element of judgment involved in our assessment of how well a theory fits observed facts; but that part of this assessment which can be quantified surely should be, and the likelihood of the theoretical values seem very well suited for this purpose.

If one theory gives predicted values $t_1, t_2, t_3, \cdots$, for a series of parameters $\theta_1, \theta_2, \theta_3, \cdots$, independently determined, while another theory predicts values $t'_1, t'_2, t'_3, \cdots$, then the first theory gives the better fit if

$$L(t_1)L(t_2)L(t_3) \cdots > L(t'_1)L(t'_2)L(t'_3) \cdots, \tag{13}$$

and vice versa if the inequality is reversed. For reasons just indicated, in addition to other complications which may arise, it may well be difficult, and not altogether necessary, to actually carry through such a comparison with full numerical precision. But the principle is there, and rough estimates can be made. Low values of L for particular parameters will indicate in what respects a proposed theory needs modification.

It is in a specific scientific situation such as we are contemplating that the difficulties of the subjective Bayesian position appear most strongly. For, quite apart from the fact that anyone who could make a plausible survey of the prior distributions involved in particle physics would have achieved a major feat of scientific imagination, there is the question, what would a posterior distribution do for us which the likelihood function will not? We could, of course, place bets rationally; but this is not a serious aim—even though it may perhaps be indulged in for amusement. And if we found a theory T to be more probable, a posteriori, than a theory T', while being less likely on the data, this could only be because T was considered a priori much more plausible than T'.

But on what grounds could we go against the evidence in this way? On examination, any case where such a thing appeared to be happening would surely turn out to be one in which we were taking account of further relevant information, in the guise of the prior distribution; however, then the likelihood computation, taking this further information into account, would agree with the posterior probability. In introducing a prior distribution, we thus are abandoning the important claim to objectivity in return for an illusory gain.

It is an important advantage of the likelihood method that, when the number of observations grows large enough, we can appeal to asymptotic theory to derive arguments of the significance test type, which enable us to rule out certain theories as too much at variance with the facts to be worth serious consideration. But the probability levels we would invoke would normally be much smaller than those commonly used, for example, with confidence intervals. For instance, the probability of decay of a Λ particle into a proton, a muon, and a neutrino, has been estimated to be about 1.5×10^{-4}, so that we might, by over-simple application of statistical arguments, be led to suppose that a particle seen to decay in this way was not a Λ particle.

Among the set of 95% confidence intervals we might obtain for the (perhaps) 1000 quantities independently measured in particle physics, there should be some 50 which would fail to contain the true values. We would have no indication of which 50 these might be; and it could easily happen that one theory gave values inside all the confidence intervals, whereas another gave several values outside their confidence intervals. Yet the latter theory was more plausible, on the data, than the former.

The practical advantage of the procedure suggested—to specify the values at which $g(\theta)/g'(\hat{\theta})$ equals 0, ± 2, and ± 4—arises from the fact that, in combining the results from two independent experiments we merely have to add the corresponding g's—easily done graphically—while in going out as far as $g(\theta) = \pm 4$, we would be taking care of any departures from linearity in $g(\theta)$ which would be likely to arise in practice.

5. Singularities in the likelihood function

Some writers have felt that the use of likelihood by itself cannot be justified in general, because infinite values may be encountered in connection with hypotheses which are not at all plausible. For example, if we have a single observation x from a normal distribution with unknown mean and unknown variance, the likelihood function is proportional to

$$L(\mu, \sigma) = (1/\sqrt{2\pi}\sigma) \exp -\tfrac{1}{2}((x - \mu)/\sigma)^2, \tag{14}$$

which becomes infinite when $\mu = x$ and $\sigma = 0$. Or again, as Bruce Hill has pointed out, when we fit the three-parameter log-normal distribution, $y = \log (x - \tau)$, with y normal with mean μ and standard deviation σ, there is a singularity as τ approaches the smallest of the observations x_i. And I am indebted to Dr. A. W. F. Edwards for drawing my attention to a most interesting example, arising in the theory of evolutionary genetics, where we suppose we are observing a Brownian movement in x with variance increasing at unit rate per unit time t. Given two pairs of observations (x_1, t_1), (x_2, t_2), with $t_1 < t_2$, we want to estimate the epoch t_0, and the value x_0 of x, at which the process began. We find a singularity of the likelihood at $(x_0, t_0) = (x_1, t_1)$.

It is trivial that all these paradoxes would be disposed of if it were accepted that the distributions involved were never really continuous, but always discrete, and that the continuous expressions for them were approximations introduced for mathematical convenience. For then the likelihood function, like the discrete probabilities, would necessarily always be finite. And it is an important fact of nature that the resolving power of any measuring instrument is always finite, so that the distribution of any quantity really observable is always discrete. But it may be felt that the quantitative effect of this discretization will be insufficient—although the likelihood will cease to be infinite, it will remain larger than intuition would lead us to expect. Therefore, some further discussion is called for.

Let us take first the (obviously imaginary) case of the single observation from a normal population. Bearing in mind the discretization, together with the important point that likelihood measures *relative* plausibility, this says that if we have an arbitrarily precise observation x, this provides arbitrarily strong evidence in favor of an hypothesis H which says that x will certainly have this value, as against any hypothesis H' which gives x a nonzero variance. Now let us imagine an experimental situation which might fancifully be thought to correspond—one in which a space probe is sent to Mars to measure its magnetic field, if any. The hypothesis H asserts that Mars has no magnetic field whatsoever, and H' asserts that there is a weak magnetic field which, like that of the earth, fluctuates with time, continuously. We suppose the fluctuation to be such that if measured at a randomly chosen epoch, such as that of the arrival of the probe, the magnetic field will be normally distributed, with mean 0.01 and standard deviation 0.02. The probe takes one reading at its nearest approach to Mars, and then begins transmitting the binary digits of the magnetic field strength. These are all zero. To begin with, we shall feel that there is little evidence in favor of either H or H', but surely, as the successive digits (assumed to be utterly reliable) turn out to be zero, we shall feel more and more disposed to favor H as against H'. In fact, as the digits accumulate, without limit, so our disposition to favor H against H' will increase without limit, just as the likelihood function suggests it should.

To turn now to the case of the log-normal distribution on analysis, it turns out that the distribution of x, which gives rise to the singularity here, is one which has a very high but narrow peak at the lowest observed value of x, x_0, say, together with a long, low tail covering larger values of x. Thus the hypothesis to which the likelihood is pointing is one which says, in effect, that a moderate fraction (perhaps 20%) of the values of x will lie very near to x_0, that no value of x will be observed less than x_0, while the remaining (say) 80% of values will be spread out rather thinly over the range above x_0.

Now the data discussed by Hill related to the date of appearance of symptoms after inoculation for smallpox, and were in fact recorded to the nearest day. On carrying through the arithmetic, it appears that the paradox could only arise in this case if we supposed the times of appearance of symptoms to be recorded to the nearest $\exp\{-150\}$-th part of a day, that is, correct to about 10^{-59} seconds. It may perhaps bring home the absurdity involved to point out that merely to avoid the uncertainty due to the finite velocity of light signals, it would be necessary, for this sort of accuracy, for the distance from the observer to the subject to be controlled to less than the nuclear radius. But now suppose we had a medical theory which predicted the time of occurrence of symptoms with this sort of precision; would not observations in accordance with it be taken as strong confirmation?

6. Likelihood sets and confidence intervals

If the view is accepted that the likelihood serves to rank possible parameter values in a rational order of plausibility on given data, then those subsets $A(\lambda;\, x)$ of the parameter space defined by

$$A(\lambda;\, x) = \{\theta\colon L(\theta, x) \geq x\} \tag{21}$$

acquire some importance. For a given x, those points belonging, for example, to $A(\tfrac{1}{2};\, x)$ will be such that no alternative value of θ can be specified for which the given data will give a likelihood ratio more than 2 to 1 in favor. In this sense, these are the values of θ which this experiment will not by itself contradict. Or, we may convert the values of λ, ranging from 0 to 1, to another scale of value of C, also ranging from 0 to 1, by the condition

$$\int_{A(\lambda, x)} L(\theta, x)\, dm \Big/ \int_{\Omega} L(\theta, x)\, dm = C, \tag{22}$$

and so obtain a measure, C, of the assurance with which we might assert that θ was in $A(\lambda;\, x)$. This measure of assurance would have the frequency interpretation referred to in the above section; it would not, in general, be a "confidence coefficient," in the sense of confidence interval theory.

Finally, we might establish λ as a function of x, with parameter α, in such a way that the resulting $A(\lambda;\, x)$ would be a confidence set for θ, with confidence coefficient α. In this way we can obtain three ways of parametrizing the sets A, for a given x, with the parameters λ, C, and α, all of which will range between 0 and 1. The parametrization with λ will involve nothing which is not given in the original specification; the parametrization with C will involve an arbitrary choice of the measure m for its interpretation; the parametrization with α will also involve an arbitrary classification of the observed result along with others not observed. In the asymptotic situation to which the classical maximum likelihood theory applies, natural choices in the two latter cases will lead to an equation of C with α, and a unique functional relationship of each with λ. In small samples, this will not be so; and perhaps the more liberal minded of us will feel inclined to quote all three parametrizations, since all of them convey useful information.

REFERENCES

[1] L. J. SAVAGE, "Subjective probability and statistical practice," *Foundations of Statistical Inference*, London, Methuen, 1962.

[2] FRANK SOLMITZ, "Analysis of experiments in particle physics," *Ann. Rev. Nuclear Science*, Vol. 14 (1964), pp. 375–402.

[3] R. A. FISHER, *Statistical Methods for Research Workers*, London, Oliver and Boyd, p. 10, 1958 (13th ed.).