Statistics and Computing (1997) **7**, 247–252

The direct use of likelihood for significance testing

A. P. DEMPSTER

Department of Statistics, Harvard University, USA.
E-mail: dempster@stat.harvard.edu

An approach to significance testing by the direct interpretation of likelihood is defined, developed and distinguished from the traditional forms of tail-area testing and Bayesian testing. The emphasis is on conceptual issues. Some theoretical aspects of the new approach are sketched in the two cases of simple vs. simple hypotheses and simple vs. composite hypotheses.

Keywords: hypothesis test, likelihood inference, postdictive interpretation, predictive interpretation.

1. Introduction

The importance of likelihood is recognized in widely differing approaches to inference. This paper is concerned with some natural but underdeveloped uses of likelihood which have interesting theoretical consequences and are potentially useful for statistical practice. The body of the paper develops a specific use of likelihoods as a substitute for tail-areas in traditional significance testing, following the heuristic proposition that, if a null hypothesis H_1 has likelihood 1/20 or less of the likelihood of an alternative hypothesis H_2, then H_1 can be rejected with much the same force as a null hypothesis rejected on the basis of a tail-area of size 1/20 or less.

In order to motivate the general attitude taken here towards the concept of likelihood, consider the following pair of precepts for a working statistician making use of the tools of statistical inference. First, he should keep clearly in mind the distinction between probabilistic devices which yield directly interpretable probabilities or expectations and those which deal only tangentially with a particular situation. Second, he should recognize and identify in his work a mixture of forward and backward modes of operation, where forward operation means the use of adopted probability models for arriving, perhaps tentatively, at estimates, predictions or decisions, while backward operation means testing, evaluating and revising the probability models in use. It will be argued below that the direct interpretation of likelihood falls more naturally into backward operation than into forward operation.

The contrast between direct and indirect roles in inference is plainly visible in the context of likelihood. Likelihood is fundamental to frequentist inference because the likelihood function is a sufficient statistic and thence is associated with a broad array of optimality properties. But frequentist inference deals only with operating characteristics of decision procedures which are directly relevant to an ensemble of usually hypothetical data sets including the data set under analysis, and hence frequentist inference is direct inference only by covert misinterpretation. Similarly, likelihood is fundamental to Bayesian inference, but in the indirect technical role of a multiplier which converts prior odds to posterior odds.

R.A. Fisher discussed indirect aspects of likelihood, especially in relation to efficiency, sufficiency and exhaustive estimation, and in relation to the fiducial argument. But Fisher also recognized direct consideration of the likelihood function as a variety of inference. As an illustration of how to interpret the likelihood function directly, he exhibited a range of parameter values with likelihoods at least 1/15 of the maximum of the likelihood. His verbalization of the meaning of the interval was:

Reprinted, with kind permission of the author and editor of the original proceedings, from Memoirs No. 1, Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, 7–22 May 1973, pp. 335–54. (eds. O. Barndorff-Nielsen, Preben Blaesild, and Geert Schou), Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

> Values of the parameter outside the last limits are obviously open to grave suspicion. (Fisher, 1958)

What is the connection between likelihood and suspicion which Fisher regarded as obvious? If it is maintained that likelihood is a primitive concept to be interpreted separately from probability, then the connection with suspicion is scarcely obvious, but if the connection is channeled through probability it becomes reasonably transparent. The first step is to note that likelihood values are probabilities assigned *a priori* to an outcome subsequently found to have occurred, or, in the words of the original definition:

> Likelihood: The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed. (Fisher, 1922)

The connection between probability and suspicion is that the occurrence of an event of small probability raises doubts, the familiar example being the ordinary significance test whose interpretation was phrased by Fisher as follows:

> The force with which such a conclusion is supported is logically that of the simple disjunction: *either* an exceptionally rare event has occurred, *or* the theory of random distributions is not true. (Fisher, 1958)

When Fisher cautioned against confusing likelihood with probability, he was stressing that the implied retrospective interpretation of probabilities should not be confused with the prospective interpretation which would be in order if the likelihood function could be justified as a fiducial or Bayes posterior density function.

I believe that retrospective interpretation is much more used in applied inference than is prospective interpretation. It is rarely discussed openly, perhaps because the existence of doubts or suspicions implies prior beliefs. It is not popular to admit that probabilities relate to beliefs, even though it is difficult to understand inference about particular data sets otherwise. A troublesome difficulty regarding retrospective interpretation is embodied in the question: which among the generally large number of improbable events that occurred shall be singled out for retrospective interpretation? A restriction to likelihood, if acceptable, would go far to answer that question. In an earlier attempt to focus attention on the contrast between the two forms of direct inference (Dempster, 1964), I introduced the terms postdictive and predictive, feeling then as now that a distinctive new term was needed to dramatize the retrospective mode.

The position just sketched relates likelihood primarily to the backward or testing phase of inference, rather than to the forward or estimating phase, and therefore runs counter to the dominant view that likelihood is primarily a tool for estimation. The latter view can easily be documented in Fisher's works, which here as elsewhere appear as something less than a logically unified whole, but the contradiction is not deep because most of Fisher's earlier writing on likelihood concerned indirect technical roles in the theory of estimation, and his work on the fiducial argument was largely an attempt, I believe unsuccessful, to define conditions under which postdictive probabilities could legitimately be converted to predictive probabilities. His direct interpretation of likelihood functions, apart from the controversial fiducial interpretation, was in tune with the more modest postdictive line of interpretation which I seek to develop in this paper.

An important school of statistical inference, represented most recently by Edwards (1972), suggests direct interpretation of likelihood as a measure of support for various hypotheses. I believe, however, that the retrospective nature of likelihood interpretation clearly invalidates the positive connotations of the term support. Likelihood values should be regarded as measuring *fit to* hypotheses rather than *support for* hypotheses. The former claims less, and avoids the charge that one is attempting to gain by stealth and confusion the interpretation which rightly belongs mainly to Bayes.

After a brief discussion of simple vs. simple hypothesis testing, the paper addresses subtler issues associated with composite hypotheses. Bayesian posterior distributions across composite hypotheses will be introduced as appropriate tools, which in turn will lead to a central thesis of the paper, namely that Bayesian *posterior distributions of likelihood* are important practical tools for assessing the fit of data to hypotheses.

2. Testing a simple null hypothesis against a simple alternative hypothesis

Although rare in practice, the simple vs. simple case defines a base point for the study and comparison of different versions of significance testing. In this section, tail-area testing and Bayesian testing will be described first, for reference, and then likelihood testing will be defined and related to the two traditional versions. The notation provides that an observable X has density $f_1(x)$ under the null hypothesis H_1 and $f_2(x)$ under the alternative hypothesis H_2.

The three versions of testing share a common form of rule for rejecting H_1, namely, that H_1 is rejected if

$$L(X) = f_2(X)/f_1(X) \geq c \tag{2.1}$$

for some pre-chosen level c. The differences lie in the rationale behind the choice of c.

Under tail-area testing, the choice is based on considerations of size

$$P_1(c) = \Pr(L(X) \geq c | H_1) \tag{2.2}$$

and power

$$P_2(c) = \Pr(L(X) \geq c|H_2). \tag{2.3}$$

If size is present to a small conventional level such as 0.05 or 0.01, then the postdictive suspicion interpretation may be associated with a rejection of H_1. The celebrated Neyman–Pearson fundamental lemma provides the tail-area theory with a justification for the form (2.1) by showing that likelihood ratio testing rules have maximum power for given size.

Bayesian testing relies on posterior probabilities $Q_1(x)$ and $Q_2(x) = 1 - Q_1(x)$ of H_1 and H_2, which in turn are determined by prior probabilities Q_1 and $Q_2 = 1 - Q_1$ of H_1 and H_2 together with the observed $L(X) = f_2(X)/f_1(X)$, according to the formula

$$Q_1(X) = Q_1/(Q_1 + Q_2 L(X)). \tag{2.4}$$

The principle is to reject H_1 if $Q_1(X)$ falls to a conventional level α, such as $\alpha = 0.05$ or 0.01, which principle leads automatically to a rule of the form (2.1) with

$$c = \frac{1-\alpha}{\alpha} \Big/ \frac{1-Q_1}{Q_1}. \tag{2.5}$$

As with tail-area testing, to accept H_1 in the face of the rule would be to accept the correctness of an assertion tagged with a small probability, in this case a conditional probability which remains predictively valid given the data.

Finally, the third version of testing, which I call *likelihood testing*, chooses c directly at a conventional level such as 20 or 100. The argument is the direct postdictive interpretation which I ascribed above to Fisher, namely that, if the likelihood $f_2(X)$ under the alternative hypothesis is at least c times the likelihood $f_1(X)$ under the null hypothesis for reasonably large c, then acceptance of H_1 necessarily implies that a relatively improbable outcome must have occurred, which in turn raises doubts about the possible correctness of the null hypothesis.

Tail-area testing and Bayesian testing are quite different in concept, and may in practice yield sharply different judgements about accepting or rejecting a null hypothesis. These differences lie close to the centre of the long-standing debates over the correct foundations for statistical inference. After contemplating examples where sharply differing results appear, many statisticians are led to strongly prefer one approach or the other, but there is today no unanimous judgement. The debate is complicated by the fact that the Bayesian can obtain any value at all for c in (2.5) by adjusting the prior probability Q_1.

Likelihood testing is in various ways a buffer between tail-area and Bayesian testing. Bayesians will seek to claim likelihood testing as their stepchild, approving its conformity with the likelihood principle, and noting that the Bayesian who adopts a conventional $\alpha = 0.05$ and an indifference prior probability $Q_1 = 0.5$ operates with $c = 19$ which is very close to the $c = 20$ which the 0.05-minded likelihood tester would be inclined to adopt. Such a Bayesian would be wrong, however, for the essence of likelihood testing, as defined here, is postdictive interpretation which is conceptually free of any involvement with prior probabilities for H_1 and H_2. The operational similarity between likelihood testing and a particular plausible rule of Bayesian testing is a happy accident in the sense of making it possible to debate the difference between tail-area testing and the shared likelihood–Bayes rule of testing in a conceptual framework which is not confounded with the awkward issue of prior probabilities.

Once this perspective is established, it is but a short step to recognizing that there is an elementary but fundamental mathematical connection between tail-area testing and likelihood testing in the simple vs. simple case. Compare the following with Birnbaum (1969), p. 129.

Lemma. The size $P_1(c)$ and power $P_2(c)$ of the tail-area likelihood ratio test (2.1) satisfy the inequality

$$P_1(c)/P_2(c) \leq 1/c, \tag{2.6}$$

and hence $P_1(c)$ alone satisfies the weaker inequality

$$P_1(c) \leq 1/c. \tag{2.7}$$

Proof. Suppose that R denotes the region in the space of X values satisfying (2.1), and that S denotes the subregion of R where $f_1(x) \neq 0$. Then

$$\begin{aligned} P_1(c) &= \Pr(f_2(X)/f_1(X) \geq c|H_1) \\ &= \int_R (1) f_1(x)\, \mathrm{d}x = \int_S (1) f_1(x)\, \mathrm{d}x \\ &\leq \int_S \frac{f_2(x)}{f_1(x)} \frac{1}{c} f_1(x)\, \mathrm{d}x = \frac{1}{c} \int_S (1) f_2(x)\, \mathrm{d}x \\ &\leq \frac{1}{c} \int_R (1) f_2(x)\, \mathrm{d}x = \frac{1}{c} P_2(c), \end{aligned}$$

as required. □

As an example, suppose that an observation X^* is just significant according to the tail-area test of size 0.05 and power 0.80, so that the critical constant c for this test is expressible as $c = f_2(X^*)/f_1(X^*)$. The inequality (2.6) then asserts that $f_2(X^*)/f_1(X^*) \leq 16$ while (2.7) asserts that $f_2(X^*)/f_1(X^*) \leq 20$. Thus, if a direct likelihood test with $c = 20$ and tail-area test of size 0.05 are considered nominally equivalent, the likelihood test is more conservative in the sense of rejecting less often. The interpretation holds for general α, not simply $\alpha = 0.05$.

The inequality allows for a large measure of conservatism in favour of likelihood testing. Consider, for example, an artificial situation where X is a single real-valued observation on (0,1) and $f_1(x) = 1$ on $0 < x < 1$. Suppose that

$f_2(x)$ is positive and increasing on $(0, 1)$, so that rejection regions of the form (2.1) can be equivalently expressed in the form

$$X \geq 1 - \alpha \tag{2.8}$$

Where α is the size of the test considered as a tail-area test. Select α to be any conventional small value, such as $\alpha = 0.01$. It is then easy to select $f_2(x)$ so that $f_2(1 - \alpha)$ is arbitrarily small, such as 0.02, while $\int_{1-\alpha}^{1} f_2(x)\mathrm{d}x$ is arbitrarily close to 1, say 0.99. The result is a tail-area test with arbitrarily small size and arbitrarily high power, and yet with the seemingly paradoxical property that an observation that is just extreme enough to reject has an associated likelihood ratio highly favoring the null hypothesis. For example, a tail-area test at size 0.01 and power 0.99 may reject while the likelihood factor stands at 50:1 favoring the null hypothesis.

Extreme examples like the foregoing are good intellectual devices for the re-examination of basic principles. I believe that various arguments, such as the censoring argument of Pratt (1962), weight the principles strongly in favor of total conditioning as is embodied in likelihood testing, and hence I find the rail-area test results invalid in the foregoing example, and thence untrustworthy in general. On the other hand, specific hypotheses like $f_1(X)$ and $f_2(X)$ are themselves very often untrustworthy in practice. The result is that neither approach should be ruled out in all cases, and that a choice should be made depending on individual circumstances. Tail-area testing will often be judged more robust, which may make up for weakness in principle. Whichever way, it is clearly desirable to have a choice of approaches available.

3. Testing a simple null hypothesis against a composite alternative hypothesis

When H_2 is a composite hypothesis parametrized by $\theta \epsilon \Omega_2$, the likelihood ratio statistic becomes

$$L(X; \theta) = f_2(X; \theta)/f_1(X) \tag{3.1}$$

depending on the unknown value of the parameter θ, where $f_1(x)$ is as above the density of the observable X under the simple null hypothesis H_1, and $f_2(x; \theta)$ denotes the family of densities of X which specify the composite hypothesis H_2 as θ ranges over Ω_2. Since θ is unknown, direct consideration of the likelihood ratio as in the simple vs. simple case is not possible without the introduction of a special device. The device proposed here depends on the availability of an acceptable posterior distribution of θ given the observed X, which in turn implies a posterior distribution of $L(X; \theta)$. Likelihood testing thus comes to involve the direct interpretation of the posterior distribution of $L(X; \theta)$ for a fixed observed X.

Although alternative approaches to posterior distributions exist, such as the fiducial argument or its various relatives, the Bayesian approach will be assumed here. However, since the requirement is for a posterior distribution only within H_2, the Bayesian analysis requires only a prior density $q_2(\theta)$ over Ω_2 and does not involve prior probabilities Q_1 and $Q_2 = 1 - Q_1$ over H_1 and H_2.

A fully Bayesian analysis would of course be possible, based on posterior probabilities $Q_1(X)$ and $Q_2(X) = 1 - Q_1(X)$ computed according to the formula

$$Q_1(X) = Q_1/(Q_1 + Q_2 \int L(X; \theta) q_2(\theta)\mathrm{d}\theta). \tag{3.2}$$

The main reason for choosing to avoid dependence on Q_1 and Q_2 is that testing is conceived here as treating the null and alternative hypotheses asymmetrically. The null hypothesis has preferred status in the sense of being accepted unless the evidence in the data rules it out. There are dangers of misintepretation in this concept of 'accept', and these must be recognized openly, as in all cases of postdictive data interpretation. Within the asymmetric framework, however, it cuts across the grain to suppose an even-handed allocation of prior probabilities Q_1 and Q_2 and to allow this allocation the key role implied by (3.2).

Once the backward phase of inference is completed, and supposing H_2 to have been accepted, then the requirement of a prior density $q_2(\theta)$ is often seen as necessary to forward uses of the model for prediction or decision-making. If this is so, then some involvement of $q_2(\theta)$ in the testing procedure is natural. Two alternative forms of involvement will now be defined and contrasted.

A comparison of (2.4) and (3.2) suggest that an obvious extension of likelihood testing from the case of simple H_2 to composite H_2 would be to replace direct interpretation of $L(X)$ with direct interpretation of

$$L^*(X) = \int L(X; \theta) q_2(\theta)\ \mathrm{d}\theta. \tag{3.3}$$

In support of this view, it may be noted that

$$L^*(X) = h_2(X)/f_1(X) \tag{3.4}$$

where

$$h_2(X) = \int f_2(X; \theta) q_2(\theta)\ \mathrm{d}\theta. \tag{3.5}$$

The form of (3.4) is identical to that of (2.4) except that $f_2(X)$ is replaced by $h_2(X)$, the latter being the marginal density of X under the Bayesian version of H_2.

The case against direct assessment of $L^*(X)$ may be illustrated by considering the effect of a change of scale in the prior density $q_2(\theta)$. If θ ranges over ordinary k-dimensional space Ω_2, then a change from $q_2(\theta)$ to $b^{-k} q_2(\theta/b)$ typically introduces a factor of roughly b^{-k} into $h_2(X)$, and thence into $L^*(X)$, while at the same time the posterior density over Ω_2 changes little. The phenomenon emerges clearly as b increases, since it typically happens

that $b^k h_2(X)$ tends in the limits as $b \to \infty$ to the integral of the likelihood function, while the posterior density of θ converges to the normalized likelihood function. In other words, $L^*(X)$ is unstable against increasing the scale of $q_2(\theta)$ while the posterior distribution of θ stabilizes. The resulting stable posterior inferences may or may not conform adequately with posterior inferences from realistic prior assessments, but at least these posterior inferences are generally commensurate. By contrast, whatever the value of X, a comparison of $h_2(X)$ and $f_1(X)$ will show the weight piling up on $f_1(X)$ as b increases, effectively ruling out the possibility of rejecting H_1. When b is large, the marginal prior assessment on X is typically distorted in an implausible way which results in small $h_2(X)$. Thus, a small value of $L^*(X)$ is mainly a reflection of the implausibility of the scaled prior density $b^{-k}q_2(\theta/b)$, and does not imply that certain members of the family $f_2(X;\theta)$ are not in fact much more plausible than $f_1(X)$.

The recommended approach is to reject H_1 if the ratio $f_2(X;\theta)/f_1(X)$ is reasonably sure to be large given knowledge of X, where weight is put on values of the unknown θ in proportion to their plausibility given X. For example, one might ask to be 0.90 sure that $f_2(X;\theta)/f_1(X) \geq 100$. A general version of likelihood testing can be described as the (γ, c) rule, where H_1 is rejected if the posterior probability that $f_2(X;\theta)/f_1(X) \geq c$ is at least γ. The computation of γ derives from the posterior distribution of θ over Ω_2, which depends on the prior distribution $q_2(\theta)$ but typically not in a highly sensitive way. In practice, the critical constants γ and c are to be selected on heuristic grounds.

Although conceptually quite different from tail-area testing, the (γ, c) version of likelihood testing will often produce similar results except that an element of conservatism creeps into the rejection process. To see this, consider the following canonical example:

$$\begin{aligned} X &= (X_1, X_2, \ldots, X_k) \\ \theta &= (\theta_1, \theta_2, \ldots, \theta_k) \end{aligned} \tag{3.6}$$

where $X_1, X_2, \ldots, X_k$ are independent with $\mathrm{N}(0,1)$ distributions under H_1 and are independent with $\mathrm{N}(\theta_i, 1)$ distributions under H_2. The example is canonical because a wide class of k-parameter models can be reduced to approximately this form in large samples. The associated canonical form of prior distribution is the improper density $q_2(\theta) \equiv 1$ which can be acceptable as an adequate approximation to many more realistic prior assessments, especially in the case of applications to large sample estimation. As is well known, the posterior distribution of θ associated with $q_2(\theta) \equiv 1$ asserts that $\theta_1, \theta_2, \ldots, \theta_k$ are independently distributed with $\mathrm{N}(X_i, 1)$ distributions. Also, it is easily checked in the example that

$$L(X;\theta) = \exp\left(-\frac{1}{2}\sum_{i=1}^{k}(X_i - \theta_i)^2 + \frac{1}{2}\sum_{i=1}^{k}X_i^2\right). \tag{3.7}$$

Putting these two facts together, it follows that the posterior distribution of

$$-2 \log L(X;\theta) = \sum_{i=1}^{k}(X_i - \theta_i)^2 - \sum_{i=1}^{k}X_i^2 \tag{3.8}$$

is an ordinary χ^2 on k degrees of freedom but shifted to the left by $\sum X_i^2$. It follows that the (γ, c) likelihood test as defined above rejects H_1 in favour of H_2 provided that

$$\sum_{i=1}^{k}X_i^2 \geq \chi^2(k,r) + 2\log c \tag{3.9}$$

where $\chi^2(k,r)$ denotes the r quantile of the χ^2 distribution on k degrees of freedom.

Since $\sum X_i^2$ is the χ^2 statistic of ordinary tail-area testing it is apparent from (3.9) that the $(\gamma, 1)$ test is formally identical to the standard χ^2 test at level $1 - \gamma$. Thus, if γ is close to unity, the standard test can be reinterpreted from the likelihood testing standpoint as saying that the statistician is nearly certain *a posteriori* that the unknown $L(X;\theta)$ is at least unity. To the likelihood tester, knowledge that $L(X;\theta) \geq 1$ is not in itself any evidence that H_1 is surprising relative to H_2. Given an observation which has borderline significance according to the χ^2 test, the likelihood tester can say he is nearly sure that H_2 fits better than H_1 according to the likelihood ratio criterion, but to reject he must ask for the extra measure of evidence provided by $2 \log c$.

In practice, the likelihood tester might be willing to back off on γ, say to 0.80 or 0.60, while maintaining c at standard levels such as 20 or 100. Under this scheme, the likelihood test becomes less conservative relative to the χ^2 test as k increases, and even rejects more often for large k, as is clear because the $2 \log c$ criterion is free of k while the standard deviation of the χ_k^2 distribution increases like $\sqrt{k}$. The interchange appears at first sight to invalidate the general observation that likelihood tests are conservative, but the appearance may be deceiving because a Bayesian would be increasingly suspicious of the uniform prior element $\mathrm{d}\theta_1, \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_k$ as k increased. A more reasonable Bayesian or empirical-Bayesian prior element would almost certainly cluster more closely about H_1, thus returning an element of conservatism to the likelihood test. As the number of parameters increases, the premise that reasonable inference can be done without gambling on representations of genuine prior knowledge of parameters becomes increasingly shaky, and a plausible Bayesian view would be that the tester should welcome a theoretical framework which straightforwardly allows for what is, in effect, a necessity.

4. Concluding remarks

The foregoing sections are intended to lay a groundwork for the study of a range of commonly adopted hypotheses.

In real world examples, the null hypothesis is almost always composite, as is the alternative hypothesis, so that posterior distributions of likelihood must be found both over H_1 and over H_2. The correct way to compare these posteriors needs careful study.

The emphasis above has been on situations where the null hypothesis H_1 is essentially a point in a higher dimensional parameter space H_2. This situation generalizes to the more common model-building decision problem of whether or not to add parameters, where H_1 is a manifold in the space H_2. Two other testing situations are also important, which may be described as series and parallel hypothesis pairs. An example of the former would be $H_1 : \mu \leq 0$ vs. $H_2 : \mu > 0$, the general case being a separation of one large space into two simply connected parts. An example of the latter would be $H_1 : X \sim \mathrm{N}(\mu, \sigma^2)$ for $-\infty < \mu < \infty, 0 < \sigma < \infty$ vs. $H_2 : X \sim \mathrm{C}(\xi, \eta^2)$ for $-\infty < \xi < \infty, 0 < \eta < \infty$, where N and C denote location and scale families based on standard normal and standard Cauchy distributions, respectively. In general, parallel H_1 and H_2 would be non-intersecting families of the same dimension and roughly similar modelling intent. Testing series hypotheses occurs mainly in the context of testing a scientific hypothesis rather than testing the adequacy of fit of some speculative parametric form. Testing parallel hypotheses could be part of a modelling operation, as in the example where the choice of a family is preparatory to drawing inferences, say about the centre of a symmetric univariate population. Parallel hypotheses could also be scientifically interesting, as in the case of comparing two functional forms of a non-linear regression model.

In the case of series or parallel testing, it may often be inappropriate to select out one hypothesis as the null hypothesis to be accepted unless disproven. In such circumstances it would be of interest to compare posterior distribution of likelihood under matched pairs of hypotheses, and observe, for example, that one distribution is scaled up or down relative to the other by a factor of 2 or 5. The indication would be that one is fitting better than the other, not enough to disprove one or the other, but enough to suggest that a tentative modelling decision had better go one way rather than the other.

Finally, it should be emphasized here that any use of postdictive reasoning, whether from tail-areas or from likelihoods, involves a deliberate choice of a tenuous and less-than-ideal form of logic. Especially because the results can differ sharply from Bayesian inferences, they need careful and separate interpretation. It is accepted in this paper that significance testing has a legitimate place in the collection of inference tools. But it is entirely beyond the scope of the paper to try to define the limits on this use, especially *vis a vis* Bayesian alternatives.

References

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

Dempster, A. P. (1964) On the difficulties inherent in Fisher's fiducial argument. *Journal of the American Statistical Association*, **59**, 56–66.

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

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* A, **222**, 309–68.

Fisher, R. A. (1958) *Statistical methods and scientific inference*, 2nd edn. Oliver and Boyd, Edinburgh.