Likelihood*

by

David V. Hinkley
University of Minnesota
School of Statistics
Technical Report No. 376
July 1980

*Research supported by NSF Grant MCS-7904558

LIKELIHOOD*

David V. Hinkley**

University of Minnesota, School of Statistics

July 1980

* This is an expanded version of a Special Invited Lecture given at the IMS Western Region Conference, in Davis, California on June 16, 1980.

**Research supported by NSF Grant MCS-7904558.

1. Introduction

To a large extent the present paper gives only a brief personal overview of some recent and ongoing research into likelihood and its role in statistical inference. Certainly there is no attempt to present a comprehensive review. For the most part the details of particular results are omitted, since such details are available in the very recent literature.

Some background is presented briefly in the next two sections, dealing respectively with models and inferential purpose (Section 2) and basic likelihood theory including information recovery (Section 3). Section 4 describes very recent research on large-sample likelihood inference for single parameter models, and the effect of goodness-of-fit tests is discussed in Section 5. An incomplete view of multiparameter problems (Section 6) is followed by some general remarks and brief references to important aspects of likelihood inference not treated in the main discussion.

2. Statistical Models and Likelihood Inference

2.1 On Models

A common starting point for many statistical analyses is the specification of a statistical model, that is a probability model with relevant physical characteristics imbedded as parameters. In some cases one may make the model larger that is thought necessary for purposes of testing a physical theory. The specification of the probability model is not a trivial step, even though common usage of the phrase "let $X_1,\ldots,X_n$ be i.i.d. with p.d.f. ..." suggests otherwise.

The most tangible forms of statistical models are those involving a real act of random sampling - either in selection of observation units

from an extant finite population, or in selection of a design layout determining treatment allocations to experimental units in a comparative experiment. In the former case well-defined parameters exist which are, in principle, determinate; whereas in the comparative design problem this is not generally so. In both cases probability is a well-defined consequence of randomization. Measurement-error problems are perhaps next-most tangible, in that we believe indefinite repetition of the measuring process is possible, in principle, and that the measured quantity is essentially determinate. But at this point we begin to face the fact that many statistical models involve hypothetical random sampling. It is then more difficult, yet equally necessary, to define precisely what is meant by saying that "data $x_1,\ldots,x_n$ is as a random sample from a population wherein x has probability distribution $F(x,\theta)$." I would contend that in some instances this statement involves an abstraction not unlike that involved in stating a prior distribution for θ. It would then seem to me to be important to recognize that the notion of long-run repeated sampling is a fiction of convenience, unless it is a physical fact.

This latter view might seem more valid when our modelling is a consequence of preliminary data-screening, which is often the case... Statistical analysis is, after all, adaptive because model analysis is contingent on the correctness of the model.

Debates over the various meanings of statistical probability have, of course, helped to make statistics a lively subject for centuries. One particularly instructive debate was that between Fisher and Jeffreys (see Lane, 1980). Recent evidence that the old issues never die may be found in Basu (1980) and discussion thereof.

2.2 Likelihood and Inferential Purpose

Once a statistical model has been formulated, various statistical operations can be performed - tests of fit, tests of special-case hypotheses, estimations, predicitons, etc. The majority of these involve likelihood, which may be defined loosely as the relative probability with respect to the parameter of an outcome arbitrarily close it the observation. If our model specifies the necessary series of probability densities (continuous, discrete or both) for successive individual observations $x_1, x_2, \ldots,$ and if we observe $x_1, \ldots, x_n$, then we say that the likelihood at θ given $x_1, x_2, \ldots, x_n$, denoted by $\text{Lik}(\theta | x_1, \ldots, x_n)$, is

$$f_1(x_1|\theta) \prod_{j=2}^{n} f_j(x_j | x_1, \ldots, x_{j-1}, \theta) \ .$$

We shall ignore inessential, though interesting, technical difficulties associated with Likelihood definition. For the majority of the discussion we will assume that $x_1, \ldots, x_n$ are independent outcomes, so that with $\underset{\sim}{x}_n = (x_1, \ldots, x_n)$

$$\text{Lik}(\theta | \underset{\sim}{x}_n) = \prod_{j=1}^{n} f_j(x_j|\theta) \tag{2.1}$$

whose logarithm is denoted by $\ell_n(\theta)$ or $\ell(\theta | \underset{\sim}{x}_n)$. The special case $f_j \equiv f$ will also be assumed here, this involving no appreciable loss of generality in terms of what follows.

Our general purpose is two-fold: first, to reduce $\underset{\sim}{x}_n$ to one or a few statistics that are essentially equivalent to $\underset{\sim}{x}_n$ for all practical purposes. Second, to construct a meaningful probability statement involving θ and an estimate thereof without unavoidable sacrifice of

"information." These steps would be accomplished by computing a statistic $s(x_1,\ldots,x_n)$, which may be an estimator, and then finding a quantity $Q\{s(x_1,\ldots,x_n);\theta\}$ with known probability distribution in an appropriate reference set of sample outcomes.

My own view of this may be phrased somewhat differently as follows. We desire a fully informative probability statement that is capable of combination with a prior distribution $p(\theta)$ when such is specified, but which is also capable of generating confidence intervals when no $p(\theta)$ is specified. The first capability is possessed by $\text{Lik}(\theta|\underset{\sim}{x}_n)$ without reference to its repeated sampling properties, but this is not enough for the second capability. (A theorem of A. Birnbaum (1962) claims that two familiar principles of repeated-sampling inference imply that $\text{Lik}(\theta|\underset{\sim}{x}_n)$ contains all the "evidence", but this does not tell us what the inferential method is.)

The evident lack of precision in our formulation of purpose is, regrettably, unavoidable because approximations are involved. I have phrased the formulation is this way because it seems to me entirely legitimate to ask how much, and what, the data tell us before we apply Bayes's Theorem, even for a rigid Bayesian.

3. Some Basic Likelihood Theory

At this juncture three severe restrictions will be placed on our model so as to simplify the discussion: (i) θ is one-dimensional, (ii) n is (treated as) fixed, and (iii) the likelihood is regular. Both (i) and (ii) will be removed later, and (iii) is regarded as practically unimportant.

The Likelihood function (2.1) defines the minimal sufficient statistic $\underset{\sim}{s}$ for θ. Unfortunately, unless $\{f(x|\theta)\}$ is a complete exponential family, $\underset{\sim}{s}$ has larger dimension of variation than θ. Therefore, in general, reduction of s to an estimator (single statistic) will involve loss of information. For convenience and familiarity, information will be measured here by Fisher's method, namely by defining

$$\text{information in statistic } T = \mathcal{I}_\theta^{(T)} = E\{-\ddot{\ell}(\theta|T)\} \quad ;$$

each $\cdot$ denotes one differentiation with respect to θ.

Particular interest focusses on the MLP (maximum likelihood point) $\hat{\theta}$, a solution of $\dot{\ell}(\theta|s) = 0$, for which*

$$\mathcal{I}_\theta^{(s)} - \mathcal{I}_\theta^{(\hat{\theta})} > 0 \qquad (3.1)$$

outside the complete exponential families. Fisher was justifiably concerned with this result. The following simple example illustrates why.

Let e_1, e_2 be independent Bernoulli outcomes, such that $e_i = \pm 1$ with equal probability, and let $x_i = \theta + e_i$. The estimate

$$\tilde{\theta} = \frac{1}{2}(x_1 + x_2)$$

is such that $pr(\tilde{\theta} = \theta) = pr(e_1 + e_2 = 0) = \frac{1}{2}$. This, and the value of $\tilde{\theta}$, are all that is available if (x_1, x_2) is reduced to $\tilde{\theta}$. This sacrifices obvious information, since

*For any other statistic $\tilde{\theta}$, $(\mathcal{I}_\theta^{(s)} - \mathcal{I}_\theta^{(\hat{\theta})})/(\mathcal{I}_\theta^{(s)} - \mathcal{I}_\theta^{(\tilde{\theta})}) \sim \tilde{c} \leq 1$ as n increases where $\tilde{c}$ depends on the particular form of $\tilde{\theta}$.

$$\begin{aligned} &\text{if} \quad |x_1 - x_2| = 0 \quad , \quad \text{then} \quad \tilde{\theta} \neq \theta \quad \text{for sure} \\ &\text{if} \quad |x_1 - x_2| \neq 0 \quad , \quad \text{then} \quad \tilde{\theta} = \theta \quad \text{for sure.} \end{aligned} \tag{3.2}$$

The example, while admittedly not within the technical framework of our discussion, does illustrate a phenomenon that is - namely the existence of supplementary information in measurement-error models, discovered by Fisher (1934). As in the example, so for a general measurement-error model

$$x_j = \theta + e_j \qquad j=1,\ldots,n \tag{3.3}$$

the residuals $\underset{\sim}{a} = \{(x_j - \hat{\theta}): j=1,\ldots,n\}$ contain the supplementary information (and it is only supplementary). Fisher showed that for this model

$$\mathcal{J}_\theta^{(s)} = E[\, \mathcal{J}_\theta^{(\hat{\theta}|\underset{\sim}{a})} \,] \quad , \tag{3.4}$$

i.e. that the full sample information is "saved" in $\hat{\theta}$ together with its conditional sampling distribution given $\underset{\sim}{a}$ (this is the pair of statements (3.2) in the example). The specific algorithm presented by Fisher is remarkable: the density of the conditional sampling distribution is

$$h(\hat{\theta}|\underset{\sim}{a},\theta) = \text{lik}(\theta|\underset{\sim}{x}) / \int \text{lik}(t|\underset{\sim}{x})dt \quad , \tag{3.5}$$

Expressing this more conveniently in terms of $D = \hat{\theta} - \theta$,

$$h_D(d|\underset{\sim}{a},\theta) = \frac{\text{lik}(\hat{\theta}_{obs} - d|\underset{\sim}{x})}{\int \text{lik}(t|\underset{\sim}{x})dt} = \frac{\text{lik}(-d|\underset{\sim}{a})}{\int \text{lik}(t|\underset{\sim}{a})dt} \quad , \tag{3.6}$$

which is the supplementary information in explicit form.

The following general principle evolved from Fisher's discovery.

> Conditionality Principle: If $\underset{\sim}{S}$ contains a component $\underset{\sim}{A}$ (called ancillary statistic) whose probability distribution is independent of θ, then inference should be based on $h(\underset{\sim}{s}|\underset{\sim}{a},\theta)$ - i.e. inferential probability calculations should be conditioned on the observed value of the ancillary.

This deserves pause for comment. First, the principle guides us to an inference that is specific to the single sample outcome $\underset{\sim}{x}$, by restricting our frame of reference to samples "like $\underset{\sim}{x}$" in the manner described by $\underset{\sim}{a}$. A special reason for this will be discussed in Section 4. The essential attribute of $\underset{\sim}{a}$ is that it determines the "relevant precision" for $\hat{\theta}$; but some ancillaries do not seem to do this (Buehler, 1980).

One argument against conditioning is that if samples $\underset{\sim}{x}_n$ are really to be drawn repeatedly, then conditional hypothesis tests and confidence intervals at fixed levels will be less precise on average than unconditional counterparts. The superficiality of this criticism has been exposed by Barnard (1976, 1978). In any event, what is good on average in a genuine aggregate of samples may not correspond to what is good in a single instance - as Stein showed for simultaneous estimation.

There is (finally) an obvious practical criticism: The conditionality principle is not applicable in general, because an ancillary statistic $\underset{\sim}{a}$ is not always available. (A counter-proof of this statement has been sought unsuccessfully for 44 years!) This can be overcome by loosening the constraint of mathematical exactitude slightly. To this end, an inexact, yet practically useful, way to explain the essence of conditioning is to say that the observed $\underset{\sim}{x}$ is to be surrounded by the smallest

possible reference set in which non-degenerate sampling probability is defined, while at the same time losing the least possible sample information about θ.

The latter statement is helpful in our framework of replicated measurement, since for large n we can deal with approximations in a fairly systematic manner. Some fragmentary results concerning the first step, that of recovering information, were first given by Fisher, expanded by Rao and developed most recently by Efron (1975). Of particular interest are the results

$$\mathcal{I}_{\theta}^{(\underset{\sim}{s})} - \mathcal{I}_{\theta}^{(\hat{\theta})} = \{\gamma_{\theta}^{(\underset{\sim}{s})}\}^2 \; \mathcal{I}_{\theta}^{(\underset{\sim}{s})} + O(n^{-1}) \tag{3.7}$$

and

$$\mathcal{I}_{\theta}^{(\underset{\sim}{s})} - \mathcal{I}_{\theta}^{(\hat{\theta},-\ddot{\ell}(\hat{\theta}|\underset{\sim}{s}))} = O(n^{-1}) \quad , \tag{3.8}$$

where in (3.7) $\gamma_{\theta}^{(\underset{\sim}{s})}$ is Efron's statistical curvature* of $\ell(\theta|\underset{\sim}{s})$. The static $-\ddot{\ell}(\hat{\theta}|s)$, hereafter denoted by I, is called the observed information. The variation of I relative to $\mathcal{I}_{\theta}^{(s)}$ is essentially measured by $\{\gamma_{\theta}^{(s)}\}^2$, which helps to understand the difference between (3.7) and (3.8).

In the next section we go on to review some modern developments that make use of the "approximate recovery of information" suggested by Fisher. Section 5 returns to another aspect of the conditionality principle.

---

* $\gamma_{\theta}^{(s)}$ is the invariant $\left[\mathrm{Var}\{-\ddot{\ell}(\theta|s)\} - \mathrm{Cov}^2\{\dot{\ell}(\theta|s),\ddot{\ell}(\theta|s)\}/\mathcal{I}_{\theta}^{(s)}\right] \div (\mathcal{I}_{\theta}^{(s)})^2$.

4. New Likelihood Theory

4.1 Normal Approximation Theory

At times Fisher would refer to I as an ancillary statistic, although this is not correct in general. Note that for the measurement-error model, (3.6) shows the shape of the likelihood to be the key ancillary. The larger is n, the more nearly normal-shaped is the likelihood - and that normal shape is characterized by

$$\text{variance} = I^{-1} \quad .$$

This notion carries over to the general case, where the standarized form

$$a_2 = \frac{I - \jmath^{(s)}_{\hat\theta}}{\gamma^{(s)}_{\hat\theta}\, \jmath^{(s)}_{\hat\theta}} \tag{4.1}$$

is approximately ancillary (in fact a_2 is a proximately $N(0,1)$), and both $I(\hat\theta - \theta)^2$ and $2\{\ell(\hat\theta|s) - \ell(\theta|s)\}$ are approximately χ^2_1 given a_2. This is discussed in great detail by Efron and Hinkley (1978). It is important to recognize that $\jmath^{(s)}_{\hat\theta}(\hat\theta - \theta)^2$ behaves very differently, since for fixed a_2

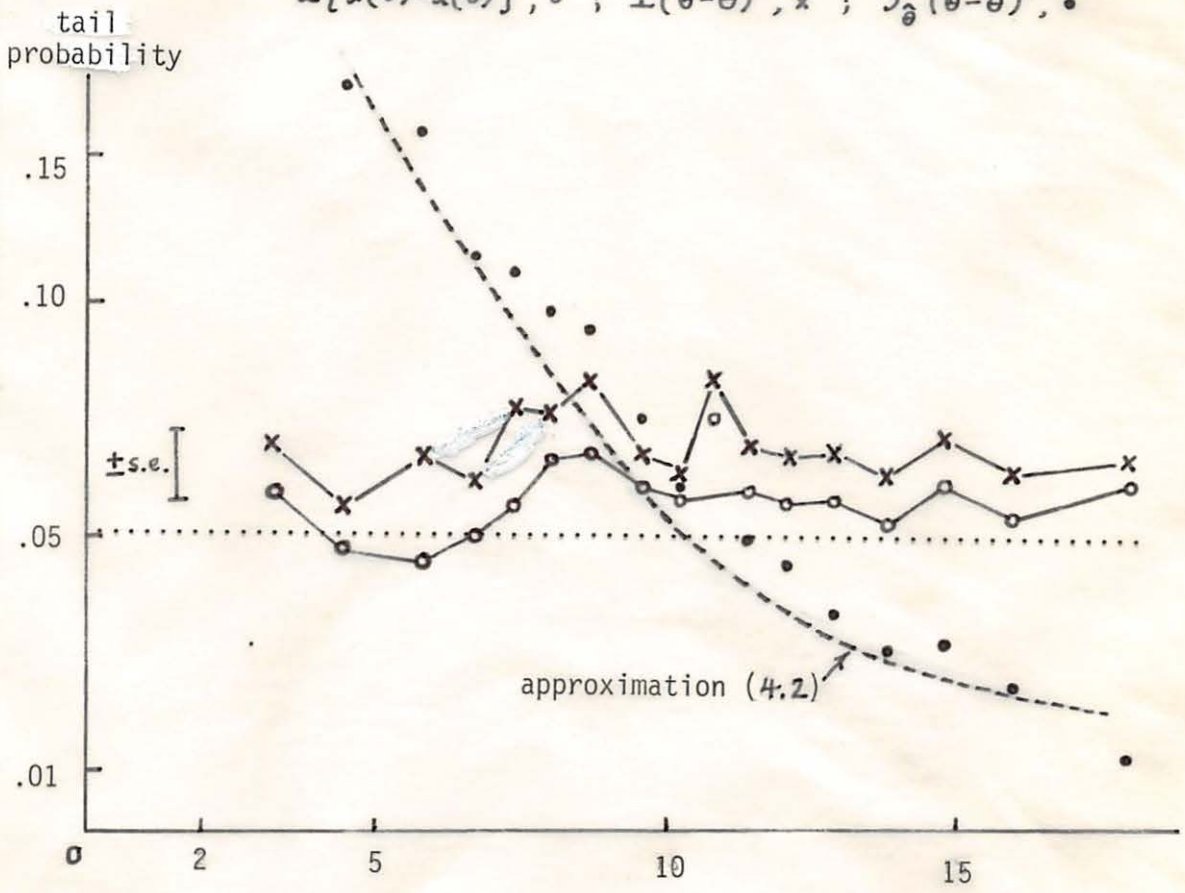
$$\jmath^{(s)}_{\hat\theta}(\hat\theta - \theta)^2 \approx (1 + a_2\gamma^{(s)}_{\hat\theta})I(\hat\theta - \theta)^2 \approx (1 + a_2\gamma^{(s)}_{\theta})\chi^2_1 \quad . \tag{4.2}$$

Figure 1 shows very reliable Monte Carlo estimates of pr(statistic $\geq$ 95% point of $\chi^2_1|a_2$) for the three statistics

$$2\{\ell(\hat\theta|s) - \ell(\theta|s)\},\ I(\hat\theta - \theta)^2,\ \jmath^{(s)}_{\hat\theta}(\hat\theta - \theta)^2 \quad ,$$

Figure 1. Tail probabilities of pivotal statistics in Cauchy measurement error model with n=20. Plotted points are Monte Carlo estimates of pr(statistic $\geq$ 3.84) for statistics

$2\{l(\hat{\theta})-l(\theta)\}$, ∘ ; $I(\hat{\theta}-\theta)^2$, × ; $\mathcal{I}_{\hat{\theta}}(\hat{\theta}-\theta)^2$, •



together with the theoretical approximations .05, .05, (4.2), for the Cauchy measurement-error model when $n = 20$.

4.2 More Accurate Theory

The normal approximation for $I^{1/2}(\hat{\theta} - \theta)$ is unsatisfactory in one familiar respect: the exact and approximate densities differ by a $O(n^{-1/2})$ term. Of course in the measurement-error model this can be avoided by working with the exact result (3.6). A better approximation can also be obtained in general. To see this it is convenient to work within a curved exponential family (Efron, 1975) where $f(x|\theta)$ is given by

$$f(x|\theta) = \exp(\underset{\sim}{\alpha}_\theta^T \underset{\sim}{x} + c_\theta + d_x) \quad . \tag{4.3}$$

Now $\underset{\sim}{x}$ is k-dimensional, as is $\underset{\sim}{\alpha}_\theta$, but $\{\underset{\sim}{\alpha}_\theta\}$ is a one-dimensional curve. If we write $\underset{\sim}{\beta}_\theta = E(\underset{\sim}{x}|\theta)$ and $\underset{\sim}{s} = \Sigma \underset{\sim}{x}_j$, then successive derivatives of $\ell(\theta|\underset{\sim}{s})$ at $\theta = \hat{\theta}$ are given by

$$\dot{\ell}(\hat{\theta}) = \dot{\underset{\sim}{\alpha}}_{\hat{\theta}}^T(\underset{\sim}{s} - n\underset{\sim}{\beta}_{\hat{\theta}}) \ , \ \ddot{\ell}(\hat{\theta}) = -I = -\ \mathfrak{J}_{\hat{\theta}}^{(s)} - n\ddot{\underset{\sim}{\alpha}}_{\hat{\theta}}^T(\underset{\sim}{s} - n\underset{\sim}{\beta}_{\hat{\theta}}) \ , \text{ etc.}$$

A simple affine transformation of the "shape statistics" $\ddot{\ell}(\hat{\theta}),\ldots,\ell^{(\cdot k)}(\hat{\theta})$ yields the approximate ancillary vector statistic

$$\underset{\sim}{a} = \begin{bmatrix} a_2 \\ \vdots \\ a_k \end{bmatrix} = M_{\hat{\theta}}(\underset{\sim}{s} - n\underset{\sim}{\beta}_{\hat{\theta}}) \tag{4.4}$$

which has approximate standard spherical normal density for large n . (M can be chosen so that a_j depends only on the first j derivatives

of ℓ at $\hat{\theta}$, in which case a_2 is (4.1).) Now one can obtain $h(\hat{\theta}|\theta,\underset{\sim}{a})$ via the following steps:

(i) Use of sufficiency $g(s|\theta) = g(s|\hat{\theta}(s))\exp\{\ell(\theta|s)-\ell(\hat{\theta}(s)|s)\}$

(ii) Use of an Edgeworth series expansion for $g(s|\hat{\theta}(s))$

(iii) Transformation $s \to (\hat{\theta},a)$

The result, given in Hinkley (1980), is quite simple:

$$h(\hat{\theta}|\underset{\sim}{a},\theta) = \frac{lik(\theta|\underset{\sim}{x})}{\int lik(t|\underset{\sim}{x})dt}\{1+O(n^{-1})\} \qquad (4.5)$$

<u>provided that</u> we choose θ so that $j_\theta^{(s)} \equiv n$. As in (3.6), the more convenient form for $D = \hat{\theta} - \theta$ is

$$h_D(d|\underset{\sim}{a},\theta) = lik(\hat{\theta}_{obs} - d|\underset{\sim}{x})/\int lik(t|\underset{\sim}{x})dt \quad . \qquad (4.6)$$

Thus the likelihood itself provides the appropriate approximate distribution for $\hat{\theta}$. Cox (1980) reaches the same conclusion indirectly, but more generally, by an ad hoc argument.

Note that (4.5) holds in the complete exponential case, when $\underset{\sim}{a}$ is null. The relevant expansion theroy is discussed by Barndorff-Nielsen and Cox (1979). The fact that there <u>is</u> an $O(n^{-1})$ error in (4.5) is unavoidable whenever $\underset{\sim}{a}$ is not exactly ancillary.

4.3 Sequential Estimation Experiments

Reference has been made to the observed information I as a relevant measure of the actual information content of data, or precision of estimation, as opposed to the prior expectation $j_\theta^{(s)}$. This suggests that if

one wished to obtain a specific amount of information (precision) C, say, then one should obtain data so that $I \geq C$. This implies sequential sampling of $x_1, x_2, \ldots$ until

$$N = \min\{n: I(x_1,\ldots,x_n) \geq C\} \quad . \tag{4.6}$$

Notice that if one solved the usual "design equation" $\mathcal{I}_\theta^{(s)} \geq C$ one would obtain a fixed value of n. For which the subsequent sample $(x_1,\ldots,x_n)$ may give $I << \mathcal{I}_\theta^{(s)}$; for example, with $n = 20$ in the Cauchy measurement error model we deduce from the $N(0,1)$ approximation for a_2 that I will be at least 4.5 less than $\mathcal{I}_\theta^{(s)} = 10$ with probability 10%. Of course $I >> \mathcal{I}_\theta^{(s)}$ is also possible.

The sequential scheme (3.6) has been studied in some detail by Grambsch (1980). For the measurement-error model N and I are exactly ancillary, so that (3.6) still applies. More generally, provided that θ is chosen to make $\mathcal{I}_\theta \equiv n$, a is again approximately ancillary and the earlier normal-approximation theory holds again. The superior result (4.5) probably holds also. Grambsch develops an approximate distribution theory for stopping time N, when C is large, and in particular she shows that

$$E\,N \approx C/\mathcal{I}_\theta^{(x_1)} \quad \text{and} \quad \text{var}\,N \approx C\{\gamma_\theta^{(x_1)}\}^2/\{\mathcal{I}_\theta^{(x_1)}\}$$

4.4 Peculiar Cases

The theory reviewed above rests heavily on the fact of replication and uses expansions with respect to normal approximations. Without replication rather different results can emerge. One particularly

interesting class of problems is that of regular nonergodic processes, discussed by Feigin and Reiser (1979); see also Feigin (1981). For the very special case of a Yule process $\{x_t\}$ with $x_0 = 1$, $E(x_t) = \exp(\theta t)$, the log likelihood based on $\{x_t: 0 \le t \le T\}$ is

$$\ell(\theta|x) = \text{const.} + (x_T - 1)\log\theta - (\int_0^T x_u du)\theta$$

$$= \text{const.} + S_1 \log\lambda - S_2\lambda ,$$

a curved exponential family. Here $\hat{\theta} = S_1/S_2$, $A = I/\mathfrak{I}^{(s)}_{\hat{\theta}}$ is approximately unit exponential for large T, and $(\hat{\theta} - \theta)$ is conditionally approximately $N(0, I^{-1})$ given $A = a$. The unconditional distribution of $\hat{\theta} - \theta$ is not approximately normal. The curved exponential structure will permit an expansion theory much like that leading to (3.5). The general situation for non-ergodic processes is unclear, and is worth further detailed study.

The pervasiveness of likelihood as conditional sampling distribution, as in Sections 3 and 4.2, extends to what seems to be a very different problem: Let the doubly-infinite sequence $\{x_j: j = \ldots,-1,0,1,\ldots\}$ be such that $\{x_j: j \le \theta\}$ are iid with distribution F_0 and $\{x_j: j \ge \theta + 1\}$ are iid with distribution F_1. Then $\underset{\sim}{a}$ is again the shape of $\text{lik}(\theta|\underset{\sim}{x})$ and (3.3) holds. As Cobb (1978) notes in proving this, the problem is a location (group-invariant) problem - different only in the sense that "information" about θ remains finite even with doubly-infinite sampling.

5. Goodness-of-Fit Screening and Ancillarity

Implicit in standard approaches to statistical model analysis is a condition: if there is evidence that the given model does not fit, then

one will not base an analysis on an assumption of the model's validity. Should we not therefore take explicit account of this modus operandi when the model does fit? In consideration of this point, note that for the model (3.3), formal and informal tests of fit are based on the residuals $\underset{\sim}{a}$. The same is true for model (4.3), where we have in particular the chi-square test statistic

$$\Sigma(O - E)^2/\hat{var}(O) = \underset{\sim}{a}^T\underset{\sim}{a} \quad .$$

In fact all proper tests of fit are based on exact or approximate ancillaries. Suppose the statistics $t_1(\underset{\sim}{a}),\ldots,t_m(\underset{\sim}{a})$ are used with critical value t_i^o for t_i. Then clearly conditional inference is unaffected by goodness-of-fit screening, since

$$f(\hat{\theta}|\underset{\sim}{a},\theta,t_1(\underset{\sim}{a}) \leq t_1^o,\ldots,t_m(\underset{\sim}{a}) \leq t_m^o) = f(\hat{\theta}|\underset{\sim}{a},\theta) \ , \qquad (5.1)$$

However unconditional inference (not conditioning on $\underset{\sim}{a}$) is affected because

$$f(\hat{\theta}|\theta,t_1(\underset{\sim}{a}) \leq t_1^o,\ldots,t_m(\underset{\sim}{a}) \leq t_m^o) \neq f(\hat{\theta}|\theta) \quad .$$

This is a striking result in favor of conditional (and Bayesian) inference, I think.

Note that in certain cases where the t_i are explicitly defined one can evaluate the effects of screening on repeated-sampling properties of unconditional inference by approximate use of (5.1). For example, consider the normal approximation theory of Section 4.1 and suppose that $t_1(\underset{\sim}{a}) = |a_2|$ with critical value t_1^o. Assuming that both $I^{1/2}(\hat{\theta} - \theta)$ and a_2 are exactly $N(0,1)$, a simple Taylor expansion shows that

$$\text{pr}\{ j_{\hat\theta}^{1/2} (\hat\theta - \theta) \le z | t_1(\underset{\sim}{a}) \le t_1^o \}$$

$$\approx \Phi(z) - \frac{1}{2} \{\gamma_\theta^{(s)}\}^2 z\,\phi(z) \left\{ \frac{1 - 2\Phi(-t_1^o) - 2t_1^o\phi(t_1^o)}{1 - 2\Phi(-t_1^o)} \right\} .$$

If $t_1^o = z = 1.96$, for $n=10$ measurements on the Cauchy error model the correction term is approximately 0.01; or, put another way, the error of a nominal 2-1/2% one-sided test for θ would be close to 3-1/2% . The same effect would show up using a more refined distributional approximation for $I^{1/2}(\hat\theta - \theta)$.

Of course most problems involve more complicated goodness-of-fit statistics than in this example.

6. Multiparameter Problems

As is usually true, multiparameter models are considerably harder to deal with, and likelihood theory is less complete than for single parameter models. One point we must live with is the important difference between joint inference and separate inference. To see this, note that Bayes's Theorem gives, with obvious notation,

$$f_B(\theta_1,\theta_2|s) = p(\theta_1,\theta_2)\, \text{lik}\,(\theta_1,\theta_2|s) / \iint p(t_1,t_2)\, \text{lik}\,(t_1,t_2|s)\,dt_1\,dt_2$$

and hence

$$f_B(\theta_1|s) \quad \propto \int p(\theta_1,\theta_2)\, \text{lik}\,(\theta_1,\theta_2|s)\,d\theta_2 \quad . \tag{6.1}$$

No pivotal distribution for $\hat\theta_1 - \theta_1$ can be found, in general, for which combination with $p(\theta_1) = \int p(\theta_1,\theta_2)d\theta_2$ will yield (6.1), even approximately.

Thus, referring back to the "inferential purpose" of Section 2.2, we see that separate inference is precluded in general - even if only θ_1 is of interest.

To see the difficulty (and some light) from the viewpoint of the conditionality principle, consider the following example due to G. Cobb: $u_1,\ldots,u_n$ are independently sampled from $N(\lambda,1)$, and conditional on these, the n measurements x_j are independently $N(\psi u_j,1)$, $j=1,\ldots,n$. Now $s = (\Sigma u_j, \Sigma u_j^2, \Sigma u_j x_j)^T$, $\theta = (\lambda,\psi)^T$, $\hat{\psi} = \Sigma u_j x_j / \Sigma u_j^2$, $\hat{\lambda} = n^{-1}\Sigma u_j$ and $a = \Sigma u_j^2 - n\hat{\lambda}^2$ is ancillary. The "usual" practical analysis for ψ treats $u_1,\ldots,u_n$ as fixed, and conditions on $\Sigma u_j^2 = a + n\hat{\lambda}^2$ to give a $N(0,(\Sigma u_j^2)^{-1})$ reference distribution for $\hat{\psi} - \psi$. But the conditionality principle directs us to condition on a, not on Σu_j^2, and $h(\hat{\psi} - \psi | a)$ is not exactly or approximately $N(0,(\Sigma u_j^2)^{-1})$. With reference to Section 4.2, the element I^{22} of I^{-1} gives the conditional normal approximation variance for $\hat{\psi}-\psi$ given $a + n\hat{\lambda}^2$, not given a.

An interesting fact that leads to some speculation is that the two pivotal quantities of the form

$$P_1 = c_1(a)(\hat{\lambda} - \lambda), \quad P_2 = c_2(a,\hat{\lambda})(\hat{x} - x) \tag{6.2}$$

have standard normal distributions, independent of θ and a, and satisfy

$$P_1^2 + P_2^2 = (\hat{\theta} - \theta)^T I(\hat{\theta} - \theta) .$$

A remarkably similar phenomenon occurs in the measurement-error model $x_j = \mu + \sigma e_j$, for which $a_j = (x_j - \hat{\mu})/\hat{\sigma}$, $j=1,\ldots,n$. In this case, writing $\lambda = \log\sigma$, $\psi = \mu$ we find P_1 and P_2 of the form (6.2) which are approximately standard normal with correlation $r(\underset{\sim}{a})$ and satisfy

$$(P_1,P_2)\begin{pmatrix} 1 & r(a) \\ r(a) & 1 \end{pmatrix}^{-1}\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = (\hat{\theta} - \theta)^T I (\hat{\theta} - \theta)$$

which is conditionally approximately χ^2_2 ; see Hinkley (1978).

One might conjecture that for a general class of replication problems one can transform $\underset{\sim}{\theta} \to \underset{\sim}{\xi}$, $(\underset{\sim}{a},\hat{\underset{\sim}{\theta}}) \to (\underset{\sim}{a},\hat{\underset{\sim}{\xi}})$ and then obtain pivotals of the form

$$P_1 = c_1(\underset{\sim}{a})(\hat{\xi}_1 - \xi_1) \ , \quad P_j = c_j(\underset{\sim}{a},\hat{\xi}_1,\ldots,\hat{\xi}_{j-1})(\hat{\xi}_j - \xi_j) \quad j=2,\ldots,d$$

such that

(i) The P_j are approximately standard normal with correlation matrix $R(\underset{\sim}{a})$

(ii) $P^T\{R(a)\}^{-1}P = (\hat{\xi}-\xi)^T I(\hat{\xi})\,(\hat{\xi}-\xi) \approx (\hat{\theta}-\theta)^T I(\hat{\theta})(\hat{\theta}-\theta) \approx 2\{\ell(\hat{\theta}|s) - \ell(\theta|s)\}$.

This conjecture has to do with a conditional normal-approximation theory. But what about analogs of (3.5) and (4.5)? A relevant negative fact is that the matrix $\mathfrak{I}^{(s)}_{\theta}$ cannot be made independent of θ by suitable naming of θ . A very positive result for d-dimensional curved exponential families has been found by Barndorff-Nielsen (1980), who shows that if $\underset{\sim}{a}$ is similar in form to (4.4) then, with relative error of order $n^{-1/2}$,

$$h(\hat{\underset{\sim}{\theta}}|\underset{\sim}{a},\underset{\sim}{\theta}) \approx K(\underset{\sim}{a})\ |I|^{1/2} \exp\{\ell(\theta|s) - \ell(\hat{\theta}|s)\} \ ; \qquad (6.3)$$

this generalizes an intermediate result of Hinkley (1980). Grambsch (1980) shows that (6.3) implies a spherical standard normal distribution for $I^{1/2}(\hat{\theta} - \theta)$ conditional on $\underset{\sim}{a}$.

The situation is necessarily rather different for the case of infinitely-many nuisance parameters, that is to say for models where each x_j introduces

a new parameter in a pdf of the form $f(\cdot|\theta,\lambda_j)$, $j=1,\ldots n$. Here we have unreplicated information on each λ_j . One practical approach to some such problems is to define a partial likelihood (Cox, 1975) in order to eliminate the λ_j . Partial likelihood can be treated in a manner similar to that of Section 3.2, although partial likelihood does not share all of the properties of likelihood. A second practical approach is to adopt a sampling model for the λ_j , with density $p(\lambda|\alpha)$, and thus to replace the problem by giving the posterior form $f_B(\theta|s,\alpha)$ plus an inferential result based on $lik(\alpha|s)$. Yet a third, fiducial, approach is applied to the famous "weighted-means" problem by Hinkley (1979).

7. Concluding Remarks

Of necessity, this paper has a fairly narrow focus. No attempt has been made to give a general discussion of likelihood outside the framework of replicated experiments for which the information content is large. Within that framework I have not described all of the important recent work: a notable omission is the pair of papers by Bates and Watts (1980a,b) on curvature measures in non-linear regression, which from the practical standpoint is very important. Less recent, but also important, is the theoretical work of Fraser (1964), and the illuminating work of Sprott (1975). Useful thoughts about ancillarity are to be found in Kalbfleisch (1975).

In the paper I have not discussed hypothesis testing directly, nor point estimation in the decision theory sense. Relevant remarks about locally most powerful tests may be found in Efron (1975), Efron & Hinkley (1978) and Kallenberg (1981). On the subject of point estimation a valuable recent reference is the paper by Berkson (1980) and its discussion.

References

Barnard, G. A. (1976). Conditional inference is not inefficient. Scand. J. Statist., 3, 132-134.

Barnard, G. A. (1978). In contradiction to Dr. Berkson's dispraise--conditional tests can be more efficient. J. Statist. Plan. Inf.

Barndorff-Nielsen, O. (1980). Conditionality resolutions. Biometrika, 67, (to appear).

Barndorff-Nielsen, O. & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. J. R. Statist. Soc., B, 41, 279-312.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test (with Discussion). J. Amer. Statist. Assoc., 75, (to appear).

Bates, D. M. & Watts, D. G. (1980). Relative curvature measures of nonlinearity. J. R. Statist. Soc., B, 42,

Bates, D. M. & Watts, D. G. (1980). Parameter transformations for improved approximate confidence regions in nonlinear least squares. Submitted to the Annals of Statistics.

Berkson, J. (1980). Minimum Chi-square--not maximum likelihood! (with Discussion). Am. Statist., 8, 457-487.

Birnbaum, A. (1962). On the foundations of statistical inference. J. Amer. Stat. Assoc., 57, 269-326.

Buehler, R.J. (1979). Some examples of ancillary statistics and their properties. Univ. of Minnesota School of Statistics Tech. Rep. 364.

Cobb, G.W. (1978). The problem of the Nile: Conditional solution to a changepoint problem. Biometrika, 65, 243-251.

Cox, D.R. (1975). Partial likelihood. Biometrika, 62, 269-276.

Cox, D.R. (1980). Local ancillarity. Biometrika, 67, (to appear).

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency)(with discussion). Ann. Statist., 3, 1189-242.

Efron, B. & Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65, 457-82.

Feigin, P.D. (1981). Conditional exponential families and a representation theorem for asymptotic inference. Ann. Statist., 9, (to appear).

Feigin, P.D. & Reiser, B. (1979). On asymptotic ancillarity and inference for the Yule and regular nonergodic processes. Biometrika, 66,

Fisher, R.A. (1934). Two new properties of mathematical likelihood. Proc. R. Soc. A 144, 285-307.

Fraser, D.A.S. (1964). Local conditional sufficiency. J. R. Statist. Soc., B, 26, 52-62.

Grambsch, P.L. (1980). Likelihood inference. University of Minnesota Ph.D. thesis.

Hinkley, D.V. (1978). Likelihood inference about location and scale parameters. Biometrika, 65, 253-61.

Hinkley, D.V. (1979). A note on the weighted means problem. Scand. J. Statist., 6, 37-40.

Hinkley, D.V. (1980). Likelihood as approximate pivotal distribution. Biometrika, 67, (to appear).

Kalbfleisch, J. D. (1975). Sufficiency and conditionality (with Discussion). Biometrika, 62, 251-268.

Kallenberg, W. C. M. (1981). The shortcoming of locally most powerful tests in curved exponential families. Ann. Statist., 9, to appear.

Lane, D. A. (1980). Fisher, Jeffreys and the nature of probability. In R. A. Fisher: An Appreciation. Lecture Notes in Statistics, 1, Eds. S. E. Fienberg & D. V. Hinkley. New York, Springer-Verlag.

Sprott, D. A. (1975). Application of maximum likelihood methods to finite samples. Sankhya, 37 B, 259-70.