The Annals of Statistics
1979, Vol. 7, No. 4, 718–728

PREDICTIVE LIKELIHOOD[1]

BY DAVID HINKLEY

University of Minnesota

The likelihood function is the common basis of all parametric inference. However, with the exception of an ad hoc definition by Fisher, there has been no such unifying basis for prediction of future events, given past observations. This article proposes a definition of predictive likelihood which can help to remove some nonuniqueness problems in sampling-theory predictive inference, and which can produce a simple prediction analog of the Bayesian parametric result, posterior $\propto$ prior $\times$ likelihood, in many situations.

1. Introduction. In 1920 Karl Pearson [10] posed "the fundamental problem of practical statistics" as follows:

> An "event" has occurred p times out of $p + q = n$ trials, where we have no a priori knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?

Pearson's purpose was to reexamine the general applicability of Bayes's earlier solution, and the resulting controversy, described by Edwards [2], is of some interest. However, the main question seems to have been largely ignored in the intervening years, while parametric inference has dominated statistical thought.

In parametric inference, a fundamental concept is that of mathematical likelihood, on which both frequentist and Bayesian methods rest. A corresponding concept for prediction has been lacking. The present paper describes a definition of predictive likelihood developed independently by the author and Lauritzen [7], and shows how the definition relates to parametric likelihood and to Bayesian posterior predictive distributions. Most of the results are different from, and developed independently of, those in Lauritzen [7]. This paper is based on the author's special invited lecture at the 1975 Institute of Mathematical Statistics meeting in Atlanta, originally written up in [6]. A recent related reference is [9].

Outline. The basic problem is discussed in Section 2. Section 3 describes the main definition of predictive likelihood, establishes natural consistency properties and demonstrates a correspondence with Bayes posterior predictive densities. The discussion is deliberately nonmathematical and simple illustrative examples are given. Section 4 briefly deals with an extension of predictive likelihood that takes

Received April 1977; revised August 1977.

[1]The original work was supported by NSF grant MPS75-08778 and preparation of this account was supported by NSF grant MCS77-00959.

AMS 1970 *subject classifications.* 62A10, 62F25, 62A15, 60G25.

Key words and phrases. Prediction, confidence regions, exponential families, likelihood, Bayesian methods, statistical inference.

advantage of the existence of invariant ancillary statistics. Section 5 briefly reviews possible inferential uses of predictive likelihood, e.g., in defining unique frequentist prediction intervals.

Notation. Throughout the paper upper case letters denote random variables, lower case letters their realized values. The probability density, or mass, function (pdf) of any random variable X at the value x is denoted by $f(x)$; conditional pdf's are denoted $f(x|\theta)$, $f(x|y)$, etc. Special notation is used for data and predictand random variables: Y is the collection of observable random variables, and Z is the collection of predictand random variables (to be predicted after observing Y). Unless otherwise stated, S and T correspond to functions (e.g., sufficient reductions) of Y and Z respectively, and R is always the minimal sufficient reduction of all variables (Y, Z). Where Y and Z correspond to collections of individual X's, m and n respectively denote the sample sizes.

2. The problem. Suppose that $X_1, \ldots, X_{m+n}$ are random variables whose distribution is indexed by the unknown parameter θ. Our problem consists of being able to observe $Y = (X_1, \cdots, X_m)$ and wanting then to make a predictive statement about $Z = (X_{m+1}, \cdots, X_{m+n})$ based on the observation $Y = y$. The predictive statement is to be in the form of one or more confidence intervals, possibly centered on a point estimate. Let $\text{lik}(\theta|y)$ denote the likelihood function for θ given $Y = y$.

From a Bayesian point of view our problem is straightforward. Once the prior distribution of Θ is determined, and after $Y = y$ is observed, the posterior predictive distribution of Z given $Y = y$ may be calculated. To be specific, if Θ has prior pdf $p(\theta)$, then

$$f(\theta|y) \propto p(\theta)\text{lik}(\theta|y) \tag{2.1}$$

and the posterior predictive pdf of Z is

$$f(z|y) = \int f(\theta|y)f(z|y, \theta)d\theta = \frac{\int f(z|y, \theta)p(\theta)\text{lik}(\theta|y)d\theta}{\int p(\theta)\text{lik}(\theta|y)d\theta}. \tag{2.2}$$

The standard form for a Bayesian $100\beta\%$ prediction confidence region would be the highest posterior density region

$$R_\beta(y) = \{z : f(z|y) \geqslant k_\beta\} \tag{2.3}$$

where k_β is a function of y chosen so that

$$\int_{R_\beta(y)} f(z|y)dz = \beta.$$

The nonBayesian does not find our problem straightforward (standard frequentist methods are outlined in Section 5). This is in some contrast to the situation where inference about θ is required, when the notion of likelihood provides a possible parallel between Bayesian and frequentist methods. Thus we have highest posterior density Bayesian confidence regions for θ of the form

$$\{\theta : p(\theta)\text{lik}(\theta|y) \geqslant a\},$$

and likelihood-based frequentist confidence regions for θ of the form

$$\{\theta : \text{lik}(\theta|y) \geqslant b\};$$

for discussion see [1], Sections 7.2(v), 9.3(vii), 10.5(i). Both methods correspond to likelihood ordering of the θ-values when $p(\theta)$ is constant, although, of course, the confidence regions may still differ. There is no apparent parallel of (2.3) for the frequentist, because there is no apparent analog of $\text{lik}(\theta|y)$ for the predictand Z. Thus for the frequentist there is no statistical instrument for uniquely defining a prediction confidence region, whereas likelihood provides such an instrument when dealing with θ.

By analogy with (2.1), a prediction likelihood function $\text{lik}^*(z|y)$ would satisfy

$$f(z|y) = a(y, z)\text{lik}^*(z|y), \tag{2.4}$$

where $a(\cdot, \cdot)$ is determined by the marginal prior distributions of Y and Z. The remainder of the paper is devoted to defining such a function $\text{lik}^*(z|y)$ and to describing some of its properties. Thus the discussion focuses on a basis for predictive inference, rather than on methods for such inference.

3. The Lauritzen–Hinkley predictive likelihood and its properties. R. A. Fisher [5, page 134] suggested an ad hoc definition of predictive likelihood which used the notion of the degree to which values of y and z support the true hypothesis of a common θ value for the two sets of variables. We use the same basic notion, but tie it more closely to parametric likelihood. Essentially we define the predictive likelihood of the value z to be the relative frequency of the observation $Y = y$ given the value of the minimal sufficient reduction of $(Y, Z) = (X_1, \cdots, X_{m+n})$. According to this definition, what we have observed (y) becomes more likely as the predictive likelihood of the unknown (z) increases. As we shall see, the relationship to parametric likelihood is very close. Similar concepts have been used directly in the construction of frequentist confidence intervals by Faulkenberry [3] and Vit [12].

We start with a slight simplification of the problem introduced in Section 2. We suppose that $Y = (X_1, \cdots, X_m)$ and $Z = (X_{m+1}, \cdots, X_{m+n})$ are independent, but that individual components X_i need not be otherwise independent, nor need they be identically distributed. The minimal sufficient reductions of Y, Z and (Y, Z) will be denoted by S, T and R respectively; R is clearly the minimal sufficient reduction of (S, T). To emphasize that the value of r is determined by particular values s and t we sometimes write $r(s, t)$.

Because of sufficiency, prediction of Z given $Y = y$ is statistically equivalent to prediction of T given $S = s$; the conditional distribution of Z given T is completely known. We then make

DEFINITION 1. Let Y and Z be independent, with distributions indexed by the common unknown parameter θ. Let $R = r(S, T)$, S and T be the minimal sufficient reductions of (Y, Z), Y and Z respectively. Then if t is uniquely defined

by r and s, the predictive likelihood of $T = t$ given $S = s$ is

$$\text{lik}^*(t|s) = f(s|r) = f(s|r(s, t)), \tag{3.1}$$

and the predictive likelihood of $Z = z$ is

$$f(z|t)\text{lik}^*(t|s).$$

The predictive likelihood is independent of θ by sufficiency of R.

Notice that this definition is symmetric with respect to S and T, so that $\text{lik}^*(t|s) = f(t|r(s, t))$. With few exceptions (such as Example 2 below) a nontrivial result is obtained from (3.1) only in the case of exponential family random variables.

As a simple illustration of Definition 1, we use Pearson's Bernoulli problem.

EXAMPLE 1. Let $X_1, \cdots, X_{m+n}$ be independent Bernoulli variables with $\Pr(X_j = 1|\theta) = \theta$. Then if $Y = (X_1, \cdots, X_m)$ and $Z = (X_{m+1}, \cdots, X_{m+n})$, we have

$$R = \Sigma_{j=1}^{m+n} X_j = S + T, S = \Sigma_{j=1}^{m} X_j, T = \Sigma_{j=m+1}^{m+n} X_j.$$

A simple evaluation of (3.1) shows that the predictive likelihood is hypergeometric,

$$\text{lik}^*(t|s) = \binom{m}{s}\binom{n}{t} \Big/ \binom{m+n}{s+t}.$$

For the special case $n = 1$, this result gives "relative likelihoods" corresponding to the Laplace law of succession, i.e.,

$$\frac{\text{lik}^*(X_{m+1} = 1|s)}{\text{lik}^*(X_{m+1} = 0|s)} = \frac{s+1}{m-s+1};$$

this reflects a relationship between predictive likelihood and posterior predictive density established in Lemma 2 below.

There are two desirable consistency properties of predictive likelihood that are easily verified in Example 1 and that hold for general exponential families. In words, these properties are (i) predictive likelihood converges to the true density of T as the number m of observations increases; (ii) predictive likelihood converges to the parameter likelihood $\text{lik}(\theta|s)$ as the number n of variables to be predicted increases. More formally, we have

LEMMA 1. *Let $Y = (X_1, \cdots, X_m)$ and $Z = (X_{m+1}, \cdots, X_{m+n})$ be such that the X_j are independent and identically distributed with exponential family density*

$$f(x|\theta) = \exp\{-\theta x + c(\theta) + d(x)\}.$$

Let $S = \Sigma_1^m X_j$, $T = \Sigma_{m+1}^{m+n} X_j$ and $R = S + T$. Denote the maximum likelihood estimators of θ based on S alone and T alone by $\hat{\theta}_S$ and $\hat{\theta}_T$ respectively. Then, with $\text{lik}^*(t|s)$ defined by (3.1),

(i) *as $m \to \infty$ with n fixed,*

$$\text{lik}^*(t|S) = f(t|\hat{\theta}_S) + O_p(m^{-1}) \tag{3.2}$$

and

$$\text{lik}^*(t|S) = f(t|\theta) + O_p(m^{-\frac{1}{2}});$$

(ii) *as* $n \to \infty$ *with* m *fixed,*

$$\text{lik}^*(T|s) = f(x|\hat{\theta}_T) + O_p(n^{-1}) \tag{3.3}$$

and

$$\text{lik}^*(T|s) = f(s|\theta) + O_p(n^{-\frac{1}{2}}).$$

PROOF. Outlined in the Appendix.

In (3.3) the parameter θ summarizes the infinite future $(X_{m+1}, \ldots)$ and the parameter likelihood should be expected as a limiting result. It is conjectured that similar consistency results hold for the full generality of Definition 1. We might note that none of the above consistency properties hold for the definition given by Fisher [5, page 134].

Up to this point we have required that the components of Y and Z be independent, but Definition 1 can be extended to cover the case of dependent sequences of random variables. The essential points of Definition 1 are that R be minimal sufficient for (Y, Z); that S be a sufficient reduction of Y; and that R be determined by S and T where T is a function of Z. The extended definition takes account of the facts that S may need to be larger than the minimal sufficient reduction of Y, and that the minimal sufficient reduction of Z must be determined by T and S (i.e., not necessarily by T alone).

DEFINITION 2. Let R be minimal sufficient for (Y, Z). Let S be sufficient for Y and let T be a function of (Z, S) such that (i) R is determined by (S, T); and (ii) the minimal sufficient reduction of Z is determined by (S, T). In addition, we require the function $r(s, t)$ to have a unique inverse $t(r, s)$ for each value of s. Then the predictive likelihood of $T = t$ given $S = s$ is

$$\text{lik}^*(t|s) = f(s|r) = f(s|r(s, t)) \tag{3.4}$$

and the predictive likelihood of $Z = z$ is

$$f(z|s, t)\text{lik}^*(t|s). \tag{3.5}$$

The following two examples illustrate the differences between Definitions 1 and 2.

EXAMPLE 2. Let $X_1, \cdots, X_m, X_{m+1}$ be i.i.d. with uniform density on the range $(0, \theta)$, so that $Z = X_{m+1}$ is to be predicted from $Y = (X_1, \cdots, X_m)$. We define $X_{(j,j)} = \max(X_1, \cdots, X_j)$. Then the minimal sufficient statistic R is $X_{(m+1,\, m+1)}$, which is determined by

$$S = X_{(m,\, m)} \text{ and } \begin{array}{ll} T = 0 & (X_{m+1} \leqslant S), \\ T = X_{m+1} & (X_{m+1} > S). \end{array}$$

These statistics satisfy the conditions of Definition 2, and a direct calculation of (3.4) gives

$$\begin{aligned} \text{lik}^*(t|s) &= m/(m+1) && (t = 0) \\ &= ms^m/\{(m+1)t\} && (t > 0). \end{aligned}$$

It follows that the full predictive likelihood (3.5) for $Z = X_{m+1}$ is

$$\text{lik}^*(z|s) = \frac{m}{(m+1)s} \qquad (z \leqslant s)$$

$$= \frac{ms^m}{(m+1)z} \qquad (z > s).$$

Thus the likelihood is uniform over the observed range $(0, s)$.

If we had tried to apply Definition 1 with $T = X_{m+1}$, we could not have distinguished different values of x_{m+1} below s. That is, t would not be determined by s and r when $r = s$.

EXAMPLE 3. Let $\{X_j : j = 0, \pm 1, \cdots\}$ be a stationary first-order Markov binary sequence with

$$\Pr(X_{j+1} = b | X_j = a) = \theta_{ab} \qquad (a, b = 0, 1).$$

Suppose that $Y = (X_1, \cdots, X_m)$ is to be observed and that $Z = (X_{m+1}, \cdots, X_{m+n})$ is then to be predicted. For any time-connected sequence $(X_c, X_{c+1}, \cdots, X_d)$ we define the matrix of transition frequencies $\mathbf{Q}(c, d)$ by

$$Q_{ab}(c, d) = \quad \text{number of } a \to b \text{ transitions in } \quad (X_c, \cdots, X_d).$$

The minimal sufficient reduction of $(Y, Z) = (X_1, \cdots, X_{m+n})$ is $R = (X_1, \mathbf{Q}(1, m+n))$. The minimal sufficient reduction of Y must be augmented by X_m, so that

$$S = (X_1, \mathbf{Q}(1, m), X_m),$$

in order for R to be obtainable from S and Z. The necessary function of Z is

$$T = (X_{m+1}, \mathbf{Q}(m+1, m+n)),$$

which is minimal sufficient for Z. Note that S is minimal totally sufficient in the sense of Lauritzen [8].

Calculation of (3.4), although complicated, follows from results of Whittle [11]. In particular, for the special case $n = 1$ we obtain

$$\text{lik}^*(x_{m+1}|s) = \frac{q_{x_m, x_{m+1}}(1, m) + 1}{\sum_{j=0}^{1} q_{x_m, j}(1, m) + 1},$$

which corresponds to the result of Example 1 applied to row x_m of $\mathbf{Q}(1, m)$.

The augmentation of the minimal sufficient reduction of Y is clearly necessary for all Markov processes.

In Section 2 we alluded to the fact that there was no apparent analog of the factorization (2.1) for the Bayes posterior predictive density (2.2) of Z given $Y = y$. It is very easy to prove that such a factorization is possible using Definition 2.

LEMMA 2. *The Bayes posterior predictive density of T given $S = s$ factorizes as*

$$f(t|s) = \text{lik}^*(t|s)\frac{f(r(s, t))}{f(s)}, \tag{3.6}$$

where the last factor is the ratio of prior marginal densities of R and S.

PROOF. For S and T, the Bayes posterior predictive density (2.2) is

$$f(t|s) = \frac{\int f(s, t|\theta)p(\theta)d\theta}{\int f(s|\theta)p(\theta)d\theta}.$$

By the sufficiency of R, the right hand side is equal to

$$f(s, t|r)\frac{\int f(r|\theta)p(\theta)d\theta}{\int f(s|\theta)p(\theta)d\theta}$$

with $r = r(s, t)$. But, by Definition 2, $f(s, t|r(s, t)) = f(s|r(s, t)) = \text{lik}^*(t|s)$, which proves the result.

In principle, then, the only prior probabilities needed for prediction are those connected with samples of size m and $m + n$. This requires less than specification of the full prior density $p(\theta)$ in many cases; in exponential families the specification of $f(r)$ for *all* $m + n$ is equivalent to specification of $p(\theta)$.

It should be noted that the posterior predictive density is statistically equivalent to predictive likelihood only if the marginal prior density of R is constant for all admissible values of $r = r(s, t)$, rather than if the prior density $p(\theta)$ is constant. In particular, the equivalence will depend on the sampling model. For example, if in Example 1 $p(\theta) = 1$, then R has a uniform distribution. But with the same $p(\theta)$, the sufficient statistic under *inverse* sampling does *not* have a uniform distribution. This reflects the fairly obvious fact that the predictive likelihood alone does not satisfy the (strong) likelihood principle ([1, page 39]) as a basis for inference.

4. Conditional predictive likelihood. The predictive likelihood of Definition 2 gives nontrivial results only when sufficiency provides a genuine reduction of (Y, Z). Thus, for example, the definition is essentially vacuous for most location-parameter models. However, in several common statistical models where sufficiency provides inadequate reduction, the minimal sufficient statistic can be expressed in terms of an ancillary statistic (with distribution independent of θ) and a conditionally sufficient statistic of low dimension. In such cases a conditional predictive likelihood may be defined in terms of probability distributions conditional on the values of the ancillary structure, which parallels the usual conditional approach to parametric inference (see [5] and [1, page 38]). Here we briefly discuss a conditional version of Definition 1.

Suppose that Y and Z are independent, with R, S and T as in Definition 1. Let R, S and T be expressed respectively as (R^+, A), (S^+, B) and (T^+, C) where A, B and C are ancillary. Because C is ancillary with known distribution, the prediction problem is now equivalent to prediction of T^+ given C and S. Then one possible definition of conditional likelihood is as follows.

DEFINITION 3. Under the above conditions

$$\text{lik}^*(t|s) = f(c)\text{lik}^*(t^+|s, c) = f(c)\frac{f(s^+|b)f(t^+|c)}{f(r^+|a)}. \tag{4.1}$$

To avoid possible difficulties with nonuniqueness of ancillary statistics, we should limit the definition to cases where A, B and C are maximal invariants of R, S and T with respect to the group structure of the probability model. With this limitation it can be shown that $\text{lik}^*(t^+|s, c)$ is determined solely by the maximal invariant function of (s^+, t^+) together with b and c. In the case of location parameter models for i.i.d. variables, the predictive likelihood of T^+ is equal to the conditional density of the maximal invariant function of (S^+, T^+) given $B = b$ and $C = c$, so that the conditional predictive likelihood is then equal to the fiducial density of T.

Definition 3 does not in general permit a factorization of the Bayes posterior predictive density as in Lemma 2, unless B is empty.

5. Relation of predictive likelihood to frequentist prediction confidence regions. For the problem formulated at the beginning of Section 2, an unbiased $1 - \beta$ prediction confidence region for Z is a set $P_\beta(Y)$ satisfying

$$\Pr\{Z \in P_\beta(Y)|\theta\} = 1 - \beta \tag{5.1}$$

for all θ. There are two classical methods for determining $P_\beta(Y)$, to each of which the likelihood definitions of preceding sections relate.

The first method, attributed to Neyman, uses test critical regions as follows. We suppose that Y and Z have pdf's $f(y|\theta_1)$ and $f(z|\theta_2)$ respectively, and consider the hypothesis H_0: $\theta_1 = \theta_2$. If, for a specified alternative H_A, an unbiased (similar) test critical region Q_β of level β can be found, then

$$\Pr\{(Y, Z) \in Q_\beta; H_0\} \equiv \beta. \tag{5.2}$$

One region $P_\beta(y)$ satisfying (5.1) is, then, the projection of the complement Q_β^c onto the subspace $Y = y$. For the particular alternative H_A, the "natural" choice of Q_β is the uniformly or locally most powerful critical region, if such exists, and in many cases this leads to smallest confidence regions $P_\beta(y)$. *However*, the resulting confidence region is, in general, characterized by the alternative hypothesis H_A used in constructing Q_β. Thus the method as described does not uniquely define one system of regions $P_\beta(y)$, because H_A is not uniquely defined. This difficulty is removed by ordering the values of Z according to values of the predictive likelihood $\text{lik}^*(t|s)$, i.e., by requiring that

$$\inf_{z \in P_\beta(y)} \text{lik}^*(t|s) \geqslant \sup_{z \notin P_\beta(y)} \text{lik}^*(t|s). \tag{5.3}$$

This corresponds to the notion of likelihood-based confidence regions for θ ([1, page 218]).

EXAMPLE 4. Suppose that $X_1, \cdots, X_m, X_{m+1}$ are i.i.d. $N(\mu, \tau)$ with both μ and τ completely unknown, and suppose that X_{m+1} is to be predicted after observing $X_1, \cdots, X_m$. Here the possible artificial hypotheses H_0 and H_A are numerous. Definition 1 shows, after lengthy calculation, that for $m > 2$ $\text{lik}^*(x_{m+1}|\Sigma_1^m x_j, \Sigma_1^m x_j^2)$

is proportional to a Student-t density with $m - 2$ degrees of freedom (not the usual $m - 1$) for

$$\left\{\frac{m(m-2)}{(m-1)(m+1)}\right\}^{\frac{1}{2}} \frac{X_{m+1} - \bar{x}_m}{s_m},$$

where $\bar{x}_m = m^{-1}\Sigma_1^m x_j$ and $(m-1)s_m^2 = \Sigma_1^m (x_j - x_m)^2$. When $m = 2$ the predictive likelihood is proportional to

$$\{s_2^2 + 2(X_3 - \bar{x}_2)^2/3\}^{-\frac{1}{2}}.$$

We are thus led via (5.3) and (5.1) to the usual symmetric Student-t interval for X_{m+1}, i.e., using $m - 1$ degrees of freedom.

The second classical method for solving (5.1) is the pivotal method. Here a pivotal or invariant function $h(s, t)$ is used to determine a region w_β such that

$$\Pr\{h(S, T) \in w_\beta | \theta\} = 1 - \beta \tag{5.4}$$

for all θ, and then the prediction confidence region is

$$P_\beta(y) = \{z: h(s, t) \in w_\beta\}.$$

The predictive likelihood may again be used to uniquely define $P_\beta(y)$ via (5.3). Notice that in problems with group structure, to which the pivotal method is restricted, Definition 3 provides an immediate solution: $h(S, T)$ is the maximal invariant function of (S^+, T^+), whose conditional distribution determines $P_\beta(y)$. Example 4 is a trivial illustration of this.

APPENDIX

Proof of Lemma 1. The following is an outline of the proof for Lemma 1, part (i). The essential requirements are that a central limit theorem expansion hold for S and R with standardizing constant $m^{\frac{1}{2}}$, and that the maximum likelihood estimator $\hat{\theta}_S$ given $S = s$ satisfy

$$s = E(S|\hat{\theta}_s). \tag{A.1}$$

The latter is a standard property of linear exponential families. For the case stated in Lemma 1, if the X_j are i.i.d. with pdf

$$f(x|\theta) = \exp\{-\theta x + c(\theta) + d(x)\},$$

then for $S = \Sigma_{j=1}^m X_j$ we have

$$s + mc'(\hat{\theta}_s) = 0, \qquad E(S|\theta) = -mc'(\theta).$$

To obtain the first result in Lemma 1, we write

$$\text{lik}^*(t|s) = f(s|s + t) = \frac{f(s|\theta)f(t|\theta)}{f(s + t|\theta)} \qquad \text{for all} \quad \theta.$$

Then, choosing $\theta = \hat{\theta}_s$, we have

$$\text{(A.2)} \qquad \text{lik}^*(t|s) = f(t|\hat{\theta}_s) \times \frac{f(s|\hat{\theta}_s)}{f(s+t|\hat{\theta}_s)}.$$

It remains to show that the ratio in (A.2) is $1 + O_p(m^{-1})$. Now by (A.1) we can see that

$$\text{(A.3)} \qquad s + t = E(R|\hat{\theta}_s) + O(1).$$

We can then evaluate $f(s|\hat{\theta}_s)$ and $f(s+t|\hat{\theta}_s)$ using the Edgeworth expansion up to terms involving the fourth moments of X, as in [4, Section XVLI.2]; this requires the existence of the fourth derivative of $c(\theta)$. Formal substitution in the Edgeworth expansion gives

$$\text{(A.4)} \qquad f(s|\hat{\theta}_s) = \frac{1}{\sigma(m)^{\frac{1}{2}}}\phi(0)\{1 + o(m^{-1})\},$$

$$\text{(A.5)} \qquad f(s+t|\hat{\theta}_s) = \frac{1}{\sigma(m+n)^{\frac{1}{2}}}\phi\{O(m^{-\frac{1}{2}})\}\{1 + o(m^{-1})\}$$

by (A.1) and (A.3) respectively, where $\sigma^2 = \text{Var}(X|\hat{\theta}_s)$ and $\phi(\cdot)$ is the $N(0, 1)$ density. The required result follows immediately on taking the ratio of (A.4) to (A.5).

The second result in part (i) of the lemma is a consequence of an expansion of the first result using $\hat{\theta}_S - \theta = O_p(m^{-\frac{1}{2}})$. Note that in order to show weak convergence to $f(t|\theta)$ we need only assume that a central limit theorem applies to S and R so that by (A.1) and (A.3) the ratio term in (A.2) is $1 + o_p(1)$. Thus simple consistency of $\text{lik}^*(t|s)$ will hold quite generally if (A.1) is satisfied.

Part (ii) of the lemma follows from part (i) by reversing the roles of S and T.

Acknowledgments. I thank several readers of earlier versions of this paper for their help in clarifying the presentation. Particular thanks go to Steffen Lauritzen and to a very patient Associate Editor.

REFERENCES

[1] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman-Hall, London.

[2] Edwards, A. W. F. (1974). A problem in the doctrine of chances. *Proc. Conference on Foundational Questions Statist. Inference*. (Eds. O. Barndorff–Nielsen, et al.), Univ. Aarhus.

[3] Faulkenberry, G. D. (1973). A method of obtaining prediction intervals. *J. Amer. Statist. Assoc.* **68** 433–5.

[4] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*. 2nd. ed. Wiley, New York.

[5] Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. 3rd ed. Hafner, New York.

[6] Hinkley, D. V. (1976). On predictive inference. Technical report no. 254, School of Statist., Univ. Minnesota.

[7] Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models. *Scand. J. Statist.* **1** 128–134.

[8] Lauritzen, S. L. (1975). General exponential models for discrete observations. *Scand. J. Statist.* **2** 23–33.

[9] MATHIASEN, P. E. (1977). Prediction functions. Research report no. 24, Dept. Theoretical Statist., Aarhus Univ.
[10] PEARSON, K. (1920). The fundamental problem of practical statistics. *Biometrika* **13** 1–16.
[11] WHITTLE, P. (1955). Some distribution and moment formulae for the Markov chain. *J. Roy. Statist. Soc. Ser. B* **17** 235–242.
[12] VIT, P. (1973). Interval prediction for a Poisson process. *Biometrika* **60** 667–8.

DEPARTMENT OF APPLIED STATISTICS
UNIVERSITY OF MINNESOTA
ST. PAUL, MINNESOTA 55108