

## **Tail probabilities from observed likelihoods**

D.A.S. Fraser

Department of Mathematics, York University, Toronto, M3T 1P3 Canada

### **SUMMARY**

An exponential model not in standard form is fully characterized by an observed likelihood function and its first sample space derivative, up to one-one transformations of the observable variable. This property is used to modify the Lugannani and Rice (1980) tail probability approximation to make it parameterization invariant. Then, for general continuous models a version of tangent exponential model is defined, and used to derive a general tail probability approximation that uses only the observed likelihood and its first sample-space derivative. The analysis extends from density functions to distribution functions the tangent exponential model methods in Fraser (1988). A related tail probability approximation has been reported (Barndorff-Nielsen, 1988b) in the discussion to Reid (1988).

**Some keywords:** Barndorff-Nielsen's formula; Conditional inference; Differential likelihood; Exponential family; Likelihood; Saddlepoint method; Tail probabilities; Tangent model.

## 1. Introduction

The saddlepoint method (Daniels, 1954; Barndorff-Nielsen and Cox, 1979) provides extremely accurate approximations to density functions based on corresponding cumulant generating functions; an extensive review is given by Reid (1988). For a variable  $y$  with cumulant generating function  $c(t)$ , the first order approximation is

$$f(y) \approx (2\pi)^{-k/2} |\ddot{c}(\hat{\phi})|^{-1/2} \exp \{c(\hat{\phi}) - \hat{\phi}'y\} \quad (1.1)$$

where  $\hat{\phi}$  is determined by  $\dot{c}(\hat{\phi}) = y$  and  $\dot{c}$  and  $\ddot{c}$  denote first derivative vector and second derivative matrix. The usual derivation is asymptotic and provides an approximate density for  $\bar{y} = \Sigma y_i$  based on a sample  $(y_1, \dots, y_n)$  from a distribution with known cumulant generating function  $c(t)$ ; the cumulant generating function for  $\bar{y}$  is then  $nc(t/n)$ . Formula (1.1) reexpresses the approximation as a direct conversion from cumulant generating function to corresponding density function, in a sense the  $n = 1$  case.

The original method of proof involves an inversion of the characteristic function using a complex-plane path specially chosen in accord with general saddlepoint techniques. An alternative method uses an Edgeworth expansion for a tilted exponential model centered on the data point in question.

For statistical contexts the cumulant generating function arises naturally and is directly available for exponential models. For the  $k$ -dimensional case, the exponential model

$$f(y; \theta) = \exp(\theta'y - \kappa(\theta) + h(y)) \quad (1.2)$$

has cumulant generating function  $\kappa(\theta + t) - \kappa(\theta)$  for the  $\theta$  distribution; the likelihood function from a data point  $y$  is  $l(\theta; y) = \theta'y - \kappa(\theta)$  plus an arbitrary constant: for convenience, we use the term likelihood to refer generally to the logarithmic version. For the

exponential model (1.2) the saddlepoint approximation can be written

$$f(y; \theta) \approx (2\pi)^{-k/2} |j(\hat{\theta})|^{-1/2} \exp \{l(\theta; y) - l(\hat{\theta}; y)\} \quad (1.3)$$

where  $\hat{\theta} = \hat{\theta}(y)$ , and  $j(\hat{\theta}) = -(\partial^2/\partial\theta\partial\theta')l(\theta; y)|_{\hat{\theta}}$ . This approximation for the density of  $y$  can be transformed to a corresponding approximation for  $\hat{\theta}$ :

$$f(\hat{\theta}; \theta) \approx (2\pi)^{-k/2} |j(\hat{\theta})|^{1/2} \exp \{l(\theta; y) - l(\hat{\theta}; y)\} . \quad (1.4)$$

In this latter form, the expression is invariant under reparameterization:  $\theta$  need not be the canonical parameter.

As the likelihood function for a minimal sufficient statistic is given by the likelihood function of an original data array, the approximation has wide utility for exponential models. Also the approximation has been found to be extremely accurate in the general context of such models, and the numerical results in Section 5 support this.

For a general statistical model  $f(y; \theta)$  where  $\theta$  has dimension  $k$ , Barndorff-Nielsen (1983) has proposed the use of (1.4) as an approximation for the density of the maximum likelihood estimate. If the dimension of  $y$  is greater than  $k$ , then  $f(\hat{\theta}; \theta)$  needs to be interpreted as a conditional density  $f(\hat{\theta}|a; \theta)$  given some exact or approximate ancillary  $a(y)$ .

The formula is useful at an observed data point  $y^0$  because the corresponding likelihood function is typically available. The usual presentation of the formula, however, does not include a general prescription for determining the ancillary  $a(y)$ , and thus it does not lead directly to a plot of the density of  $\hat{\theta}$  for particular  $\theta$  values, unless  $\hat{\theta}$  is minimal sufficient or the ancillary  $a(y)$  is otherwise available. An affine ancillary has been suggested by Barndorff-Nielsen (1980) on asymptotic grounds. A computer implementable procedure for calculating a preferred ancillary  $a(y)$  based on differential likelihood is discussed in Fraser and Reid (1988a).

The original support for Barndorff-Nielsen's approximation is that it is exact for transformation models when renormalized, and it coincides with the saddlepoint approximation for exponential models. Some discussion and analysis of the Barndorff-Nielsen approximation may be found in Barndorff-Nielsen (1983; 1986b; 1988a, p. 213f), McCullagh (1984), Reid (1988); these use asymptotic calculations based on sample space geometry and cumulants or use transformation model theory. A nonasymptotic interpretation of the approximation using the Laplace-integral method is discussed in Fraser (1988): a transformation of  $\theta$  and of  $\hat{\theta}$  is defined to yield constant observed information and an approximating exponential model then supports a local saddlepoint calculation.

Lugannani and Rice (1980), also Daniels (1987), use saddlepoint methods to directly approximate a distribution function or tail probability formula. For a real variable  $y$  with cumulant generating function  $c(t)$  and distribution function  $F(y)$ , the approximation can be written

$$F(\hat{\theta}; \theta) = F(y) \approx \Phi(z) + \phi(z) \left( \frac{1}{z} - \frac{1}{\zeta} \right) \quad (1.5)$$

where  $\phi$  and  $\Phi$  are the standard normal density and distribution functions, and  $\dot{c}(\hat{\phi}) = y$  defines  $\hat{\phi}$  for  $z = \text{sgn}(\hat{\phi})[2\{\hat{\phi}y - c(\hat{\phi})\}]^{\frac{1}{2}}$ , and  $\zeta = \hat{\phi}\{\ddot{c}(\hat{\phi})\}^{\frac{1}{2}}$ ; the notation  $F(\hat{\theta}; \theta)$  allows the use of this expression later in the paper, in particular in connection with (1.6) and (1.7) below. For statistical contexts the formula is well suited to exponential models where the cumulant generating function is available from likelihood. For such models the formula is generally understood to be extremely accurate, better than integrating the approximate density (1.3) or (1.4) unless exact when renormalized. For the exponential model (1.2) in the real variable case,  $z$  becomes the signed square root of the likelihood-ratio statistic,  $\zeta$  becomes the standardized maximum likelihood

estimate, and (1.5) records the left tail probability  $F(y) = F(\hat{\theta}; \theta)$ :

$$z = \text{sgn}(\hat{\theta} - \theta)[2\{l(\hat{\theta}; y) - l(\theta; y)\}]^{\frac{1}{2}} \quad (1.6)$$

$$\zeta = (\hat{\theta} - \theta)|j(\hat{\theta})|^{\frac{1}{2}}, \quad (1.7)$$

In Section 2 we show how an exponential model not in standard form can be fully characterized by an observed likelihood function and its first sample-space derivative, up to one-one transformations of the observable variable elsewhere on the sample space. This result is used in Section 3 to modify the Lugannani and Rice formula so that it is independent of the parameterization of the model. Then for a continuous statistical model we derive in Section 4 an approximating or tangent exponential model at a point  $y$  and then use the modified Lugannani and Rice formula to obtain a general model version of that tail probability approximation. In Section 5 we discuss briefly the choice (Fraser and Reid, 1988a) of direction for differentiating the likelihood function and then illustrate the closeness of approximation using several examples and model types.

## 2. How observed likelihood determines an exponential model

Consider a continuous statistical model  $f(x; \phi)$ , that is an ordinary  $k$ -dimension exponential linear model in terms of some one-one equivalent canonical variable  $y(x)$  and one-one equivalent canonical parameter  $\theta(\phi)$ . Often when the functional form of  $f(x; \phi)$  is available, simple manipulation of the logarithm will be enough to obtain the model in canonical form. In this section we develop a procedure that uses only an observed likelihood function and its first sample space derivative at a data point  $x_0$  and produces directly the cumulant generating function, the canonical parameter, and the local canonical variable; in effect, the procedure gives a characterization of an

exponential model in terms of likelihood properties local on the sample space. This is then used in Sections 3 and 4 to modify and extend the Lugannani and Rice tail probability approximation.

The general  $k$ -dimensional exponential model has probability element

$$\exp[\theta'(\phi)y(x) - \kappa\{\theta(\phi)\} + h\{y(x)\}]dy . \quad (2.1)$$

The canonical variable  $y(x)$  and parameter  $\theta(\phi)$ , cumulant generating function  $\kappa(\theta)$ , and underlying  $h(y)$  are not uniquely determined even by the full functional form  $f(x; \phi)dx$ . An affine transformation  $\tilde{y} = Cy + a$  on  $y$  requires an affine transformation  $\tilde{\theta} = C^{-1}'\theta + b$  on  $\theta$  with dual-inverse linear component; corresponding transformations for  $\kappa$  and  $h$  are  $\tilde{\kappa} = \kappa + a'C^{-1}'\theta + c + b'a$  and  $\tilde{h} = h - b'Cy + c$ . The invariant group of the canonical representations thus has parameters  $C, a, b, c$  with dimensions  $k^2, k, k, 1$ . To eliminate the indeterminacy, we standardize  $y, \theta, \kappa, h$  with respect to a sample point  $x_0$  having maximum likelihood value  $\phi_0 = \hat{\phi}(x_0)$ : we require  $y(x_0) = 0$ ,  $\theta(\phi_0) = 0$ ,  $\kappa(0) = 0$ , and then  $\partial\theta/\partial\phi'|_{\phi_0} = I$  so that the first derivative behaviour of  $\theta$  and  $\phi$  coincide at  $\phi_0$ .

Let  $l(\phi; x) = \log f(x; \phi) - \log f(x; \phi_0)$  be the likelihood function normed to the value  $\phi_0$  which is now taken as fixed, and let

$$S = S(\phi_0; x) = \frac{\partial}{\partial\phi} l(\phi; x)|_{\phi_0} \quad (2.2)$$

be the  $\phi_0$  score taken as a function of  $x$ . From (2.1) we have  $l(\phi; x) = \theta'S - \kappa(\theta)$ ; the standardization properties then give

$$\begin{aligned} \theta' &= \psi'(\phi) = \frac{\partial l(\phi; x)}{\partial S'} \Big|_{S=0} = \frac{\partial l(\phi; x)}{\partial x'} \Big|_{x_0} \frac{\partial x}{\partial S'} \Big|_{S=0} \\ &= \dot{l}(\phi; x_0) \dot{S}^{-1}(\phi_0; x_0) , \end{aligned} \quad (2.3)$$

where

$$\dot{l}(\phi; x) = \frac{\partial}{\partial x'} l(\phi; x) \quad (2.4)$$

$$\dot{S}(\phi; x) = \frac{\partial}{\partial x'} S(\phi; x) = \frac{\partial}{\partial \phi} \dot{l}(\phi; x) \quad (2.5)$$

are *sample space* derivatives, and

$$\kappa(\theta) = -l(\psi^{-1}(\theta); x_0); \quad (2.6)$$

we call  $\theta$  the natural parameter. Note that the natural parameter  $\theta$  and the cumulant generating function  $\kappa(\theta)$  are defined entirely in terms of  $l(\phi; x_0)$  and  $\dot{l}(\phi; x_0)$ . A function  $k(x) = \partial S(\phi; x) / \partial x' |_{\phi = \hat{\phi}(x)}$ , closely related to  $\dot{S}(\phi; x)$ , has been used for variable change  $x \longleftrightarrow \hat{\phi}(x)$  in Fraser and Reid (1988a).

The exponential model  $f(x; \phi)$  can then be written as

$$\begin{aligned} f(x; \phi) &= f(x; \phi_0) \exp[\theta' S - \kappa(\theta)] \\ &= g(S; \phi) |\dot{S}(\phi_0; x)| \end{aligned} \quad (2.7)$$

and the corresponding model for the score as

$$g(S; \phi) = g(S; \phi_0) \exp[\theta' S - \kappa(\theta)]; \quad (2.8)$$

The standardization properties, for example  $\partial \theta / \partial \phi' |_{\phi_0} = I$ , are easily verified.

The likelihood  $l(\phi; x_0)$  and its sample space derivative  $\dot{l}(\phi; x_0)$  determine  $\theta$  by (2.3) and  $\kappa(\theta)$  by (2.6) and thus fully determine the exponential model (2.8) for the score variable  $S$ . For computation,  $\theta$  and  $\kappa(\theta)$  are of course directly available but the null density  $g(S; \phi_0)$  would generally require a Fourier inversion from  $\kappa(\theta)$ ; for statistical purposes, however, an accurate approximation for  $g(S; \phi_0)$  is directly available by the

saddlepoint method.

The likelihood  $l(\phi; x_0)$  and its derivative  $\dot{l}(\phi; x_0)$  also determine (2.7) the density  $f(x_0; \phi)$  of the original variable  $x$  at  $x_0$ . The density  $f(x; \phi)$  elsewhere on the sample space for  $x$  requires  $|\dot{S}(\phi_0; x)|$  and would not be available from the  $x_0$  likelihood information.

We complete this section by recording results for the more general case where  $x$  is not minimal sufficient. For this, consider a continuous model  $f(x; \phi)$  where the parameter  $\phi$  and the minimal sufficient statistic have dimension  $k$  but the variable  $x$  has dimension  $n$ . Let

$$l(\phi; x) = \log f(x; \phi) - \log f(x; \phi_0) \quad (2.9)$$

be likelihood normed with respect to some fixed value  $\phi_0$ .

The minimal sufficient statistic locally at the point  $x$  (for example, Fraser 1966) is given by

$$dl(\phi; x) = \sum_{i=1}^n \frac{\partial}{\partial x_i} l(\phi; x) dx_i = \sum_{i=1}^n l_i(\phi; x) dx_i \quad (2.10)$$

and takes values in  $L_{MS}\{l_1(\cdot; x), \dots, l_n(\cdot; x)\}$ , which is a  $k$ -dimensional vector space of  $\phi$ -functions. Let  $V = V(x) = (v_1, \dots, v_k)$  be  $k$  linearly independent vectors. We assume that the directional derivatives

$$\frac{dl(\phi; x)}{dV} = \left\{ \frac{dl(\phi; x)}{dv_1}, \dots, \frac{dl(\phi; x)}{dv_k} \right\}$$

span the space  $L_{MS}$ . This would happen typically, unless a chosen  $v_i$  fell in the  $n - k$  dimensional null space of the linear forms (2.10). Formulas (2.4) and (2.5) can then be written

$$\dot{l}(\phi; x) = \frac{d}{dV} l(\phi; x) \quad (2.11)$$

$$\dot{S}(\phi; x) = \frac{d}{dV} S(\phi; x) = \frac{\partial}{\partial \phi} l(\phi; x). \quad (2.12)$$

For the general result we now assume that the continuous model  $f(x; \phi)$  is exponential with a  $k$ -dimensional canonical parameter  $\theta = \psi(\phi)$  that is one-one equivalent to  $\phi$ , but somehow the exponential structure is disguised. We also assume that likelihood  $l(\phi; x_0)$  has been standardized with respect to  $\phi_0 = \hat{\phi}(x_0)$  and that the derivative  $\dot{l}(\phi; x_0)$  is given by (2.11). It follows then that  $l(\phi; x_0)$  and  $\dot{l}(\phi; x_0)$  uniquely determine the exponential model (2.8) for the score variable:  $\theta$  is defined by (2.3) with (2.11) and (2.12), and  $\kappa(\theta)$  is defined by (2.6).

### 3. Parameterization invariant Lugannani and Rice

The Lugannani and Rice formula (1.5) gives a left tail probability  $F(\hat{\theta}; \theta) = F(y)$  approximation for a one parameter exponential model. The accompanying definition (1.6) for  $z$  uses likelihood drop from  $\hat{\theta}$  to  $\theta$  and is invariant under reparameterization. The definition (1.7) for  $\zeta$ , however, uses the canonical parameter  $\theta$ . In this section we record a parameterization invariant version (3.1) of  $\zeta$ . The resulting modified Lugannani and Rice formula then uses (1.5) with (1.6) and (3.1).

From Section 2 we note that the canonical parameter  $\theta$  can be determined from the likelihood  $l(\phi; x_0)$  and its sample space derivative  $\dot{l}(\phi; x_0)$ , the latter being available from (2.4) or more generally from (2.11). Accordingly we obtain

$$\zeta = \{\dot{l}(\hat{\phi}; x) - \dot{l}(\phi; x)\} \dot{S}^{-1}(\hat{\phi}; x) j^{\frac{1}{2}}(\hat{\phi}). \quad (3.1)$$

In this expression we have used  $x$  for the data point and  $\hat{\phi}$  for the corresponding maximum likelihood estimate;  $\phi$  remains as the parameter value for which the tail probability

is being calculated. We also note that normalization of the likelihood function is unnecessary: that  $\log f(x; \phi)$  can be used in place of  $l(\phi; x)$ .

#### 4. General tail probability approximation

Consider a continuous statistical model  $f(x; \phi)$  with a  $k$  dimensional minimal sufficient statistic  $x$  and a  $k$  dimensional parameter  $\phi$ . The results in Section 2 show that the observed likelihood function  $l(\phi; x_0)$  and its first derivative  $\dot{l}(\phi; x_0)$  determine the model  $g(S; \phi)$  for the score variable, provided the model is exponential; they also determine the density  $f(x_0; \phi)$  at the observed data point.

Without the exponential assumption, we find it reasonable for inference to use an approximating exponential model in preference to an approximating normal model. The approximating exponential model we call a tangent exponential model although the term here differs from that in Fraser and Reid (1988a) and seems not to have direct connections with the notion of a tangent model in Amari (1987, p. 102) or with a least favourable family.

For the general continuous model  $f(x; \phi)$  we take the tangent exponential model at the point  $x_0$  to be the exponential model (2.1) that coincides with the given model at  $x_0$ ; this is given by (2.8) in terms of the score variable (2.2) described by  $S$ . This approximating model has a natural parameter  $\theta$  defined by (2.3) and has 'cumulant generating function' given by (2.6)

$$\kappa(\theta) = -l(\psi^{-1}(\theta), x_0);$$

this expression need not be a cumulant generating function but it will correspond to one, certainly to the second and perhaps to higher order in  $\theta$  about  $\theta = 0$  or in  $\phi$  about

$\phi_0 = \hat{\phi}(x_0)$ . This property was used (Fraser, 1988) to show directly that (1.4) was a general model density approximation.

Now consider the real variable, real parameter case. A left tail probability for the maximum likelihood estimate  $F(\hat{\phi}; \phi) = F(\hat{\theta}; \theta)$  is given by (1.5), with (1.6) for  $z$  and

$$\zeta = \{l(\hat{\phi}; x) - l(\phi; x)\} \dot{S}^{-1}(\hat{\phi}; x) j^{\frac{1}{2}}(\hat{\phi}) \quad (4.1)$$

for  $\zeta$ .

## 5. Application

### 5.1 Conditional inference

The approximate tail probability formulas in Sections 3 and 4 are directly applicable to a continuous statistical model with a real variable and a real parameter; they give an observed level of significance and by iteration confidence intervals.

For more general continuous model on  $R^n$  we appeal to approximate conditional inference techniques and seek a one-dimensional distribution that contains the most information concerning a real valued interest parameter  $\psi$ , and yet is as independent as possible of the remaining nuisance parameter, say,  $\lambda$ ; the full parameter  $\theta$  is thus equivalent to  $(\psi, \lambda)$ .

A one-dimensional conditional procedure on  $R^n$  is determined by an  $n - 1$  dimensional conditioning variable  $a(y)$ ; a level surface for such a variable is a one-dimensional curve. A one-dimensional curve can also be determined by its unit tangent vector  $v(y)$  at each point; we have then that  $v(y)$  satisfies the vector equation  $da(y) = 0$ ,

$$\sum_1^n \frac{\partial}{\partial y_i} a(y) v_i(y) = 0 .$$

which has  $n - 1$  coordinates.

Computer implementable methods for calculating a preferred one-dimensional conditional distribution for inference concerning  $\psi$  may be found in Fraser and Reid (1988a); the affine ancillary of Barndorff-Nielsen (1980) provides a related asymptotic approach.

Fraser and Reid (1988a) show that a vector field  $v(y)$  determines a conditional distribution

$$g(s|a; \theta) ds = k(a; \theta) f(y; \theta) \exp \left\{ \int_0^s \operatorname{div} v(y) ds' \right\} ds \quad (5.1)$$

where the variable is conveniently taken to be arc lengths, and  $\operatorname{div} v(y)$  is the divergence of the vector field. An appropriate choice of vector field  $v(y)$  and thus of a conditioning variable  $a(y)$  would make  $a(y)$  approximately ancillary for  $\psi$  and thus  $k(a; \theta)$  approximately free of  $\psi$ .

For an application of the tail probability formula to the conditional model (5.1) we need only two ingredients, the observed conditional likelihood

$$l^c(\psi; y_0) = \log f(y_0; \theta) + \log k(a_0; \theta) \quad (5.2)$$

and the observed conditional likelihood derivative

$$i^c(\psi; y_0) = \frac{d}{dv(y)} \log f(y; \theta) \Big|_{y_0};$$

interestingly, the norming constant  $k(a; \theta)$  disappears from the second formula.

Methods for obtaining approximate conditional likelihoods (5.2) are discussed in Cox and Reid (1987), Fraser and Reid (1988b). Methods for obtaining the preferred direction  $v(y)$  for differentiating the original sample space likelihood are summarized in Section 5.2 from results in Fraser and Reid (1988a).

## 5.2 Directions for differentiating likelihood

Consider a continuous statistical model  $f(y; \theta)$  where  $y$  has dimension  $n$  and  $\theta$  has dimension  $p$ . For analysis in the neighbourhood of a point  $y$  let  $\theta_0 = \hat{\theta}(y)$  and take likelihood to be  $l(\theta; y) = \log f(y; \theta) - \log f(y; \theta_0)$ . The minimal sufficient statistic at the point  $y$  is given (2.10) by  $dl(\theta; y)$  and has dimension, say  $k$ , given by that of the vector space  $L\{l_1(\cdot; y), \dots, l_n(\cdot; y)\}$ .

Conditional inference methods can be developed directly in terms of an initial variable  $y$ , which has not been reduced to minimal sufficient statistic form. For simplicity here however we assume that the reduction to the minimal sufficient statistic has been made and thus that  $y$  has dimension  $k$ . The difference  $k - p$  is the number of dimensions that the space of likelihood functions near the point  $y$  exceeds the dimension of the maximum likelihood estimate  $\hat{\theta}$ . We now summarize the methods for determining a preferred conditioning direction  $v(y)$ .

In the special case  $k = p$  the preferred direction vector  $v(y)$  satisfies

$$dU_i = 0 \quad i = 1, \dots, p - 1$$

where  $U_i = \partial l(\theta; y) / \partial \lambda_i$  is the score for the  $i$ th  $\lambda$  coordinate. These  $p - 1$  conditions can be written

$$\sum_{j=1}^k \left\{ \frac{\partial^2}{\partial \lambda_i \partial y_j} l(\theta; y_0) \right\} v_j(y_0) = 0, \quad i = 1, \dots, p - 1$$

which define  $v(y_0)$  as the vector orthogonal to  $k - 1$  vectors derived from the nuisance scores.

For the case  $k > p$  the nuisance parameter  $\lambda$  is assumed to have been orthogonalized to the parameter  $\psi$ . Let

$$S_j(\theta; y) = \frac{\partial^j}{\partial \psi^j} \log f(y; \theta) \quad j = 1, \dots, k - p$$

be the  $j$ th order score with respect to  $\psi$ ,  $I_{1j}(\theta)$  be the covariance of  $S_1(\theta; y)$  and  $S_j(\theta; y)$ , and

$$\tilde{S}_j(\theta; y) = S_j(\theta; y) - I_{1j}(\theta)I_{11}^{-1}(\theta)S_1(\theta; y) \quad j = 2, \dots, k - p$$

be the orthogonalized higher scores. Then a direction  $v(y)$  that gives likelihood change that is first order free of the nuisance  $\lambda$  and concentrates higher order  $\psi$  effects in the first order term satisfies

$$\begin{aligned} dU_i &= 0 & i &= 1, \dots, p - 1 \\ d\tilde{S}_j &= 0 & j &= 2, \dots, k - p \end{aligned}$$

Some examples showing the choice of conditioning direction  $v(y)$  are given in Fraser and Reid (1988a).

### 5.3 Numerical comparison of approximations

The conditioning technique just discussed reduces an  $n$  or  $k$  dimensional model to an approximating one-dimensional model for the interest parameter  $\psi$ ; other techniques are of course possible. We do not examine here the effectiveness of such procedures. Rather we restrict our assessment of the approximation in Sections 3 and 4 to a range of one-dimensional models; such models of course include these obtained by such conditioning procedures.

For the one-dimensional model  $f(x; \theta)$  the signed square root of the likelihood ratio statistic for testing a value  $\theta$  is

$$z = \text{sgn}(\hat{\theta} - \theta) \cdot [2\{l(\hat{\theta}; y) - l(\theta; y)\}]^{\frac{1}{2}}$$

The corresponding asymptotically-based tail probability approximation is

$$G_{LR}(\hat{\theta}; \theta) = \Phi(z) \quad (5.3)$$

for the distribution function of  $\hat{\theta}$  given  $\theta$ .

Higher order modifications to this are the Bartlett (Barndorff-Nielsen and Cox, 1984), the McCullagh (1984), and the Barndorff-Nielsen (1986a) corrections. The most recent of these corrected likelihood ratio approximations,

$$G_{CLR}(\hat{\theta}; \theta) = \Phi(z^*) \quad (5.4)$$

uses an adjusted signed-likelihood ratio statistic

$$z^* = z - z^{-1} \log K \quad (5.5)$$

which for the single real parameter case can be calculated from the approximation

$$\begin{aligned} K = & 1 + \frac{1}{6} \{ \bar{l}_3 + 3\bar{l}_{2;1} \} \bar{j}^{-\frac{3}{2}} z \\ & + \frac{1}{24} [ \{ 3\bar{l}_4 + 12(\bar{l}_{3;1} + \bar{l}_{2;2}) \} \bar{j}^{-2} + \{ 5\bar{l}_3^2 + 24\bar{l}_{2;1}(\bar{l}_3 + \bar{l}_{2;1}) \} \bar{j}^{-3} ] z^2 \end{aligned} \quad (5.6)$$

where for example  $\bar{l}_{2;1} = (\partial^2/\partial\theta^2)(\partial/\partial\hat{\theta})l(\theta; x)|_{\hat{\theta}=\theta}$  and  $\bar{j} = j(\hat{\theta})|_{\hat{\theta}=\theta}$  is the observed information evaluated at  $\hat{\theta} = \theta$ .

For the present approximation, we have

$$G_A(\hat{\theta}; \theta) = \Phi(z) + \phi(z) \left\{ \frac{1}{z} - \frac{1}{\zeta} \right\} \quad (5.7)$$

where  $\zeta$  is given by (4.1). Also, we let  $G(\hat{\theta}; \theta)$  designate the exact distribution function.

For an exponential model, the approximation (1.5, 1.6, 4.1) becomes the parameterization invariant form (1.5, 1.6, 3.1) of the Lugannani and Rice tail probability formula.

Accordingly for the first few examples we consider exponential models with  $n = 1$  : the first is the gamma model in logarithmic form with  $\theta$  as the shape parameter; the second and third are the gamma again in logarithmic form with  $\theta$  as a location parameter and two values  $p = 3, 1$  of the shape parameter. For the remaining examples we turn to models that are sometimes viewed as being ‘orthogonal’ to exponential models, certainly they can be very different: the location models  $f(x - \theta)$ . For these we progress from the gamma to the logistic and then to the long tail Cauchy distribution. The examples are summarized in Table 1.

The approximations  $G_{LR}$ ,  $G_{CLR}$ ,  $G_A$  and the exact value  $G$  for the distribution of  $\hat{\theta}$  are recorded in percent (Tables 2-7) for selected values of the basic variable  $x$ . When an  $x$  value is marked  $R$ , the complement of the distribution function, the right tail probability is recorded.

The log-gamma with shape parameter  $\theta$  (Table 2) has an exponential left tail and a doubly exponential right tail; the model is exponential in  $\theta$ . The approximations  $G_{LR}$  and  $G_{CLR}$  are high on the left and low on the right, whereas  $G_A$  coincides with Lugannani and Rice and is extremely accurate.

The log-gamma with location parameter  $\theta$  is examined for shape  $p = 3$  in Table 3 and shape  $p = 1$  in Table 4; the left tail is exponential and the right tail is double exponential. The approximations  $G_{LR}$  and  $G_{CLR}$  are low on the left and high on the right with  $G_{CLR}$  sometimes correcting in the wrong direction, whereas  $G_A$  coincides with Lugannani and Rice and is extremely accurate.

Tables 2 and 3 are describing the same distribution for  $x$  but the approximations are calculated using different exponential parameters.

The gamma with location parameter  $\theta$  is examined for shape  $p = 3$  in Table 5; the left tail is polynomial and the right tail is exponential. It is also a variable carrier

model in the parameter  $\theta$  and in this sense is extreme for the present considerations. The approximations  $G_{LR}$  and  $G_{CLR}$  are high on the left and low on the right, both well removed from the present approximation  $G_A$  and the exact  $G$ .

The logistic with location parameter  $\theta$  (Table 6) has exponential left and right tails. The approximation  $G_{LR}$  is low, whereas  $G_{CLR}$  and  $G_A$  are close and slightly high. The symmetry of the distribution may help the approximations.

As a final example we consider the extreme case of a Cauchy model with location parameter  $\theta$  (Table 7); the left and right tail are both polynomial with coefficient -2. The approximations  $G_{LR}$  and  $G_{CLR}$  are both low, at some points extremely low, whereas the approximation  $G_A$ , as a reportable observed level of significance, is acceptably close.

## 6. Discussion

A normal model has two parameters, and tail probabilities are immediately available by entering normal tables with the standardized variable. An exponential model has an infinity of parameters, and tail probabilities can be approximated with high accuracy by the Lugannani and Rice formula (1.5, 1.6, 1.7).

If a model on the real line is location normal in some parameterization, then an observed likelihood determines that model and the likelihood ratio test uses normal tables to test  $\hat{\theta}$  against a value  $\theta$ . If a model on the real line is exponential, then an observed likelihood and its first sample space derivative determine that model and a test of  $\hat{\theta}$  against  $\theta$  with saddlepoint accuracy is available using the modified Lugannani and Rice formula (1.5, 1.6, 3.1), which requires only normal tables and two entry values  $z$  and  $\zeta$ .

The underlying theme in this paper is that for more general contexts it is better to approximate at a data value using the many parameter exponential model than using just a

two parameter normal model. This argues then in favour of the approximation (1.5, 1.6, 4.1) in Section 4 in preference to the familiar likelihood ratio normal or chi square approximation. The numerical examples in Section 5 support this approach.

The limiting factor for the use of the approximation lies in the determination of the direction  $v(y^0)$  in which to calculate the derivative

$$\dot{l}(\theta; y^0) = \frac{d}{dv(y)} l(\theta; y)|_{y^0}$$

of the likelihood; this derivative is effectively the canonical parameter of the approximating exponential model. Some discussion of this may be found in Sections 5.1, 5.2.

The derivation of the approximation (1.5, 1.6, 4.1) involves two steps: a best fit by an exponential model; the asymptotic approximation to that by the Lugannani and Rice formula. This raises the question as to why the approximation should seemingly perform much better than the corrected likelihood ratio tests, say as developed by Barndorff-Nielsen (1986a), which also involve asymptotics to the same order. A partial explanation can be found in the type of derivatives used. For a real variable  $\hat{\psi}$  and parameter value  $\psi$  with  $\hat{\psi} > \psi$  the right tail probability is an integral along the strip  $\psi \times (\hat{\psi}, \infty)$  in the  $\psi \times \hat{\psi}$  plane. The present approximation uses only a first derivative for the data variable and is calculated at the start point  $(\psi, \hat{\psi})$  of the integration strip. By contrast the corrected likelihood ratio approximation (5.4) uses higher order derivatives at  $(\psi, \psi)$  removed from the integration strip. A lower order Taylor series close to the interest area would seemingly be better than a higher order series at a distant point. Formula (5.6) of Barndorff-Nielsen involves approximations for the  $K$  in (5.5); the preceding remarks would seem to apply also to the exact (5.5).

Most of the examples in Section 5 involve the location model  $f(x - \theta) = \exp \{l(x - \theta)\}$ . If the model is centred so that  $\dot{l}(0) = 0$ ,  $-\ddot{l} = \hat{j}$ , then the

present approximation (1.5, 1.6, 4.1) uses

$$\begin{aligned} z &= \text{sgn}(\zeta) \cdot [2\{l(0) - l(x - \theta)\}]^{\frac{1}{2}} \\ \zeta &= \{-\dot{l}(x - \theta)\} \hat{j}^{-\frac{1}{2}} \end{aligned} \tag{6.1}$$

For such location models, an asymptotic examination of the approximation may be found in DiCiccio, Field, Fraser (1989) together with numerical results.

More generally, consider the assessment of a real parameter in a larger location or transformation model: a simple example is given by  $\psi$  in  $f(y_1 - \psi, y_2 - \lambda)$ ; a more general example is given by the conditional analysis of say  $\beta_r$  in the nonnormal regression model  $y = X\beta + \sigma e$ . Two routes seem possible. DiCiccio, Field, Fraser (1989) use asymptotics to extend (6.1) to cover the marginal distribution of a single location variable. Fraser, Lee, Reid (1989) develop a modified conditional distribution for the location variable of interest and directly apply (6.1). The two routes lead to the same tail probability formula. In Fraser, Lee, Reid (1989) a Monte Carlo validation using statistically parallel conditional distributions is available.

### **Acknowledgements**

This is a continuation of joint work. The author also expresses appreciation for the very helpful and insightful suggestions from two referees and from the editor. Financial support is gratefully accepted from the Natural Sciences and Engineering Research Council of Canada. Special thanks go to Fissehe Abebe for computer advice, support, and computation.

## References

- Amari, S.-I. (1987). Differential geometry in statistics. Lecture Notes - Monograph Series 10, 19-94. Hayward, California: Institute of Mathematical Statistics.
- Barndorff-Nielsen, O.E. (1980). Conditionality resolutions. *Biometrika* 67, 293-310.
- Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343-65.
- Barndorff-Nielsen, O.E. (1986a). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* 73, 307-22.
- Barndorff-Nielsen, O.E. (1986b). Likelihood and observed geometries. *Ann. Statist.* 14, 856-73.
- Barndorff-Nielsen, O.E. (1988a). Parametric Statistical Models and Likelihood. Lecture Notes in Statistics 50. Berlin: Springer.
- Barndorff-Nielsen, O.E. (1988b). Discussion of paper by N. Reid. *Statist. Sci.* 3. 228-9.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and saddlepoint approximations in the statistical applications (with discussion). *J. Roy. Statist. Soc. B* 41, 279-312.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. R. Statist. Soc. B* 46, 483-95.
- Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B* 49, 1-39.
- Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* 25, 631-50.

- Daniels, H.E. (1987). Tail probability approximations. *Int. Statist. Rev.* 55, 37-48.
- DiCiccio, T.J., Field, C.A. and Fraser, D.A.S. (1989). Marginal tail probabilities and inference for real parameters, *Biometrika*, to appear.
- Fraser, D.A.S. (1966). Sufficiency for regular models. *Sankhya A* 281, 137-44.
- Fraser, D.A.S. (1988). Normed likelihood as saddlepoint approximation. *Mult. Anal.* 26,181-93.
- Fraser, D.A.S., Lee, H.S. and Reid, N. (1989). Nonnormal linear regression: an example of significance levels in high dimensions, submitted to *Biometrika*.
- Fraser, D.A.S. and Reid, N. (1988a). On conditional inference for a real parameter: a differential approach on the sample space. *Biometrika* 75, 251-64.
- Fraser, D.A.S. and Reid, N. (1988b). Adjustments to profile likelihood. *Biometrika*, to appear.
- Lugannani, R. and Rice, S.O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Prob.* 12, 475-90.
- McCullagh, P. (1984). Local sufficiency. *Biometrika* 71, 233-44.
- Reid, N. (1988). Saddlepoint methods and statistical inference. *Statist. Sci.* 3,213-38.

**Table 1.** Five models: from exponential to the Cauchy extreme

Model	Density	Model Type
log-gamma, shape	$\Gamma^{-1}(\theta) \exp \{ \theta x - e^x \}$	exponential
log-gamma, location	$\Gamma^{-1}(p) \exp \{ p(x - \theta) - e^{x-\theta} \}$	location; exponential
gamma	$\Gamma^{-1}(p)(x - \theta)^{p-1} e^{-(x-\theta)}$	location
logistic	$e^{x-\theta} (1 + e^{x-\theta})^{-2}$	location
Cauchy	$\pi^{-1} \{ 1 + (x - \theta)^2 \}^{-1}$	location

**Table 2.** Log-gamma, shape ( $\theta=3$ ): tail probability as percent; right tail when marked  $R$

x	<b>-.577</b>	<b>.423</b>	<b>1.26R</b>	<b>1.71R</b>	<b>2.14R</b>
$G_{LR}(\hat{\theta}; 3)$	2.73	23.11	28.73	7.62	.77
$G_{CLR}(\hat{\theta}; 3)$	2.40	24.96	22.52	3.87	.05
$G_A(\hat{\theta}; 3)$	1.91	19.61	31.98	8.82	.93
$G(\hat{\theta}; 3)$	1.95	19.78	31.87	8.79	.92

**Table 3.** Log-gamma, location ( $\theta=0, p = 3$ ): tail probability as percent; right tail when marked  $R$

x	<b>-2</b>	<b>-1</b>	<b>0</b>	<b>1R</b>	<b>2R</b>
$G_{LR}(\hat{\theta}; 0)$	.02	.34	5.37	56.67	3.32
$G_{CLR}(\hat{\theta}; 0)$	.01	.22	3.77	64.02	5.11
$G_A(\hat{\theta}; 0)$	.04	.63	8.05	48.92	2.21
$G(\hat{\theta}; 0)$	.04	.63	8.03	48.92	2.21

**Table 4.** Log-gamma, location ( $\theta=0, p = 1$ ): tail probability as percent; right tail when marked *R*

<b>x</b>	<b>-7</b>	<b>-3</b>	<b>-1</b>	<b>1R</b>	<b>2R</b>
$G_{LR}(\hat{\theta}; 0)$	.03	2.14	19.55	11.53	.15
$G_{CLR}(\hat{\theta}; 0)$	.07	3.43	22.51	14.44	.33
$G_A(\hat{\theta}; 0)$	.10	5.01	31.04	6.63	.06
$G(\hat{\theta}; 0)$	.09	4.86	30.78	6.60	.06

**Table 5.** Gamma ( $\theta=0, p = 3$ ): tail probability as percent; right tail when marked *R*

<b>x</b>	<b>1</b>	<b>3R</b>	<b>5R</b>	<b>7R</b>	<b>10R</b>
$G_{LR}(\hat{\theta}; 0)$	18.97	26.93	6.33	1.28	.10
$G_{CLR}(\hat{\theta}; 0)$	32.31	12.80	1.52	.11	NA
$G_A(\hat{\theta}; 0)$	7.30	43.28	12.83	3.06	.29
$G(\hat{\theta}; 0)$	8.03	42.32	12.47	2.96	.28

**Table 6.** Logistic distribution ( $\theta=0$ ): tail probability as percent.

<b>x</b>	<b>-8</b>	<b>-6</b>	<b>-4</b>	<b>-2</b>	<b>-1</b>
$G_{LR}(\hat{\theta}; 0)$	.01	.12	1.07	9.39	24.41
$G_{CLR}(\hat{\theta}; 0)$	.04	.27	1.87	12.14	27.13
$G_A(\hat{\theta}; 0)$	.04	.27	1.91	12.22	27.16
$G(\hat{\theta}; 0)$	.03	.25	1.80	11.92	26.89

**Table 7.** Cauchy distribution ( $\theta = 0$ ): tail probability as percent.

<b>x</b>	<b>-30</b>	<b>-20</b>	<b>-10</b>	<b>-5</b>	<b>-2</b>	<b>-1</b>
$G_{LR}(\hat{\theta}; 0)$	.01	.03	.12	.53	3.64	11.95
$G_{CLR}(\hat{\theta}; 0)$	.07	.15	.54	1.93	8.81	20.56
$G_A(\hat{\theta}; 0)$	.94	1.41	2.81	5.58	13.30	23.22
$G(\hat{\theta}; 0)$	1.06	1.59	3.17	6.28	14.76	25.00