A simple general formula for tail probabilities for frequentist and Bayesian inference

By D.A.S. FRASER, N. REID and J. WU

Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3

dfraser@utstat.toronto.edu, reid@utstat.toronto.edu, jian@utstat.toronto.edu

SUMMARY

We describe a simple general formula for approximating the p-value for testing a scalar parameter in the presence of nuisance parameters. The formula covers both frequentist and Bayesian contexts and does not require explicit nuisance parameterization. Implementation is discussed in terms of computer algebra packages. Examples are given and the relationship to Barndorff-Nielsen's approximation is discussed.

Some key words: Asymptotic theory; Likelihood methods; Pivotal quantities; Tail probabilities; Third order inference.

1. INTRODUCTION

We consider inference for a scalar parameter of interest $\psi = \psi(\theta)$ in a continuous statistical model with density $f(y;\theta)$, where y is a vector of length n and θ is a vector of length p. The inference is presented as an approximate p-value $p(\psi)$ for assessing an hypothesized value ψ: in the frequentist case the p-value is an approximation to the distribution function of an appropriate pivotal statistic for $\psi(\theta)$ and is described in Sections 2.1 and 3. In the Bayesian case the p-value is an approximation to the integral of the posterior marginal density for values of $\psi(\theta)$ larger than the value ψ of interest. The p-values are accurate to $O(n^{-3/2})$, do not require an explicit nuisance parametrization, and are easily calculated without recourse to special model properties.

The third order p-values will be expressed as simple modifications of familiar first order approximate p-values such as those based on the maximum likelihood and score departures

$$q_m = \frac{\hat{\psi} - \psi}{\hat{\sigma}_\psi} \ , \tag{1.1}$$

$$q_s = \ell_\psi(\hat{\theta}_\psi)\hat{\sigma}_\psi \ , \tag{1.2}$$

where $\hat{\theta}_\psi$ is the constrained maximum likelihood estimate of θ given $\psi(\theta) = \psi$, $\ell_\psi(\theta) = (\partial/\partial\psi)\ell(\theta)$ is the score function from the loglikelihood $\ell(\theta) = \ell(\theta; y) = \log f(y;\theta)$ and $\hat{\sigma}^2_\psi$ is an estimate of the variance of $\hat{\psi}$ based on observed information $\hat{\jmath}_{\theta\theta}$. First order approximate p-values are then obtained as $\Phi(q_m)$ or $\Phi(q_s)$, based on the limiting normal distribution of q_m and q_s. The estimate of variance is

$$\hat{\sigma}^2_\psi = \hat{\jmath}^{\psi\psi} = \frac{|\hat{\jmath}_{\lambda\lambda}|}{|\hat{\jmath}_{\theta\theta}|} = \psi_{\theta'}(\hat{\theta})\hat{\jmath}^{\theta\theta}\psi_\theta(\hat{\theta}) \tag{1.3}$$

where $\hat{\jmath}^{\psi\psi}$ is the p,p entry in $(\hat{\jmath}_{\theta\theta})^{-1}$ and $\psi_\theta(\theta) = \partial\psi(\theta)/\partial\theta$ is the gradient of $\psi(\theta)$. The final expression in (1.3) avoids the need for explicit nuisance parameterization.

A somewhat different first order p-value that tends to have better distributional accuracy is given by $p(\psi) = \Phi(r)$ where r is the signed likelihood root

$$r = \mathrm{sgn}(\hat{\psi} - \psi)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_{\psi})\}]^{\frac{1}{2}} \ . \tag{1.4}$$

We are concerned in this paper with approximations of the form

$$p(\psi) = \Phi_1(r, Q) = \Phi(r) + \phi(r)(r^{-1} - Q^{-1}) \tag{1.5}$$

$$p(\psi) = \Phi_2(r, Q) = \Phi\{r - r^{-1}\log(r/Q)\} \ . \tag{1.6}$$

The primary objective is a simple and widely applicable formula for Q that ensures the $O(n^{-3/2})$ accuracy of $p(\psi)$. The frequentist version is a generalization of (1.1); cf. (2.6) and (3.6). The Bayesian version is a generalization of (1.2); cf.(2.9) and (3.6).

There is a large literature on approximations of this type; some of it is reviewed in Reid (1996). In particular, if the dimension of the variable and the dimension of the parameter are the same, as after a reduction by sufficiency in exponential families or by conditioning in transformation models, the approximate p-values are obtained from the appropriate marginal or conditional density approximations, are accurate to $O(n^{-3/2})$ and can be expressed with r obtained from (1.4) and Q replaced by modifications of q_m and q_s. These modifications can be viewed as partial steps towards the general formula in this paper. When the dimension of the variable is larger than that of the parameter a reduction in dimension uses, either explicitly or implicitly, an approximate ancillary statistic. Barndorff-Nielsen (1986, 1991) has considered several model types where this is possible, and a discussion of approximate ancillaries for this context is given in Barndorff-Nielsen & Cox (1994, Ch. 7). Bayesian approximations to marginal posterior tail probabilities using (1.5) or (1.6) are discussed in DiCiccio & Martin (1991,1993).

No explicit ancillarity reduction is needed for approximations that have relative error $O(n^{-1})$ rather than $O(n^{-3/2})$. Several such approximations are suggested in Barndorff-Nielsen & Chamberlin (1991, 1994) and in Barndorff-Nielsen (1994). An approximation with relative error $O(n^{-1})$ that is also valid in a large deviation region is developed in Skovgaard (1996).

This paper presents a method for constructing Q to give a general formula for $p(\psi)$ that includes both the Bayesian case and the frequentist case with continuous variable of dimension greater than or equal to that of the parameter. The frequentist version generalizes the formula in Fraser & Reid (1995), and like that formula does not require specification of an approximate ancillary statistic. The Bayesian version generalizes a formula in DiCiccio & Martin (1991). In both generalizations only the parameter of interest need be specified; no explicit nuisance parametrization is needed. The approximation is easily implemented, as it requires only the observed likelihood function, the parameter of interest $\psi(\theta)$, and in the frequentist case the observed sample space gradient of likelihood based on an appropriate pivotal quantity and in the Bayesian case a prior density $\pi(\theta)$. An explicit form for an approximate ancillary statistic is not required, but is generated implicitly in the calculations through the use of the pivotal quantity which gives a set of vectors $V = (v_1, \ldots, v_p)$ described at (2.1) below. The choice of pivotal quantity is typically straightforward and natural, and is illustrated in the examples in Section 4. In many cases a sensible pivotal quantity is given by the probability integral transformation applied to each coordinate, as in Example 4.7 below.

In Section 2 we describe the formulas in Fraser & Reid (1995) and DiCiccio & Martin (1991) for the case of explicit nuisance parameterization. We also relate the first of these formulas to the r^* approximation of Barndorff-Nielsen (1991). The extension to the case without explicit nuisance parametrization is derived in Section

3. Examples are presented in Section 4.

To conclude this section we give a brief overview of the theoretical basis for the general formula: this involves a number of initiatives that at first sight seem somewhat unrelated. First, in an asymptotic model with a scalar parameter and scalar variable, the probability left of an observed data point depends only on the observed likelihood and the observed sample-space gradient of likelihood ignoring terms of $O(n^{-3/2})$. This can be demonstrated in various ways, for example, Fraser & Reid (1993), but perhaps the most transparent is by means of a Taylor expansion of log-density about the data point and its maximum likelihood value; the distribution function at the data point can then be evaluated by computer-algebra integration and all dependence on other than the observed likelihood and observed likelihood gradient disappears to the third order (Cakmak et al., 1998). Second, in the preceding context, one possible model with a given likelihood and likelihood gradient is an exponential model and such a model has the available Lugannani & Rice (1980) approximation (1.5) which gives directly a third order p-value, using for Q the standardized maximum likelihood estimate of the canonical parameter; the canonical parameter of an exponential model is available as the likelihood gradient. Third, for a p dimensional variable, p dimensional parameter and a scalar interest parameter $\psi(\theta)$, we can view the model with $\psi(\theta) = \psi$ fixed as a model with a p dimensional variable and a $p-1$ dimensional parameter. A one-dimensional statistic can be constructed, that is ancillary for the nuisance parameter to second order in a standardized neighbourhood of the observed maximum likelihood estimate $\hat{\theta}^0$ (Barndorff-Nielsen, 1986; Fraser & Reid, 1995). The distribution of this ancillary is obtained on the curve $\hat{\theta}_\psi = \hat{\theta}^0_\psi$, and the exponential model approximation above gives the p-value to third order (Fraser & Reid, 1993). This step involves the elimination of the nuisance parameter by Laplace-type integration using the p^* formula

(Barndorff-Nielsen, 1983) or its extension the tangent exponential model (Fraser & Reid, 1995). Fourth, for an n dimensional variable and a p dimensional parameter θ restricted to a standardized neighbourhood of the observed $\hat{\theta}^0$, it is shown (Fraser & Reid, 1995) that a second order order ancillary suffices for third order inference, and it is shown (Fraser & Reid, 1995, 1998) that a full dimensional pivotal quantity generates this ancillary. The preceding steps applied to the conditional model then give the third order p-value using the observed likelihood and the observed likelihood gradient in directions V tangent to the ancillary. These steps are recorded above in a historical order; in Section 2, they are described in the order used to computationally reduce a problem.

2. EXPLICIT NUISANCE PARAMETRIZATION

2.1 Frequentist case

Consider a vector response $y = (y_1, \ldots, y_n)'$ with independent scalar components and density $f(y; \theta)$. We assume that the log-likelihood function $\ell(\theta; y)$ is $O(n)$ in θ, the maximum likelihood estimate $\hat{\theta}$ converges at rate $n^{-\frac{1}{2}}$ to θ, and various differentiability properties hold as discussed for example in DiCiccio, Field & Fraser (1990). Also in this section we assume that there is an explicit nuisance parameterization $\theta' = (\lambda', \psi)$, that the dimension p of θ is fixed, and that $n \geq p$. The assumption of independence is used but that of identical distributions is not: for example each y_i could depend on μ_i which in turn is a function of θ in the manner of the link function in generalized linear models. Extensions to independence of blocks of coordinates is straightforward.

Recent approaches to third order likelihood inference involve two main steps in the dimension reduction of the variable being examined. The first step is a reduction from the dimension n of the variable y to the dimension p of the full parameter θ. The variable y may be the original response vector, or the corresponding minimal sufficient statistic: the first is typically more direct for computation but the second is more conventional. The second step is a subsequent dimension reduction of the variable from p to 1 yielding a measure of departure from what is expected under $\psi(\theta) = \psi$. We describe these in turn.

Reduction from n to p:

This reduction in dimension is obtained by conditioning on an approximate ancillary statistic, as is fairly standard in recent approaches to higher order asymptotics and is required for example to implement Barndorff-Nielsen's r^* approximation. For our purposes however it is sufficient to have the existence of an approximate ancillary, and to be able to compute the gradient of the likelihood with the ancillary statistic held fixed. We describe these briefly.

It is shown in Fraser & Reid (1995, 1998) that a second order ancillary $a(y)$ at a data point exists, and that vectors $V = (v_1, \ldots, v_p)$ tangent to the ancillary surface at the observed data can be constructed easily from a first derivative ancillary derived from a full dimensional pivotal quantity. It is shown in Skovgaard (1986) and in Fraser & Reid (1995) that a second order ancillary is sufficient to give third order accuracy for approximate p-values.

The tangent vectors V are constructed using a vector $z = (z_1, \ldots, z_n)'$ of pivotal quantities $z_i = z_i(y_i; \theta)$ that has a fixed distribution. A simple and easy choice is given by the successive distribution functions $z_i = F(y_i; \theta), i = 1, \ldots, n$, which are uniformly distributed. Other choices are illustrated in the examples in Section 4. The array V is obtained from the pivotal $z(y; \theta)$ by differentiating for fixed pivotal value,

$$V = \frac{\partial y}{\partial \theta'}\Big|_{(y^0,\hat\theta^0)} = -\Big(\frac{\partial z}{\partial y'}\Big)^{-1}\Big(\frac{\partial z}{\partial \theta'}\Big)\Big|_{(y^0,\hat\theta^0)}\;, \tag{2.1}$$

where the first expression is calculated for fixed z, and y^0 is the observed data point with maximum likelihood estimate $\hat\theta^0$.

The idea of differentiating y with respect to θ holding fixed the value of a pivotal statistic is possibly unfamiliar, and looks somewhat unusual at first glance. It is easily exemplified in the location-scale model, where the natural pivotal is $z_i = (y_i - \theta_1)/\theta_2$, and $\partial z/\partial\theta$ and $\partial z/\partial y$ are straightforward; alternatively writing $y_i = \theta_1 + \theta_2 z_i$ we obtain $\partial y/\partial\theta'\big|_{y^0,\hat\theta^0} = (1, \hat z)$ directly where $\hat z = (y_i - \hat\theta_1)/\hat\theta_2$. This pivotal is used in Examples 4.1 and 4.2 below

The construction of V was derived from the local location model proposed in Fraser (1964), and the tangent directions for the local location ancillary were then shown (Fraser & Reid, 1995) to be the same as those of a second order ancillary statistic. It is this property that enables the calculation of the third order approximation, without specific construction of the ancillary statistic.

The conditional model:

If we have an explicitly available approximate ancillary a, we can work within the conditional density, $f(t|a;\theta)$, where the dimension of t is p. From the related factorization of the full model we obtain

$$\log f(y;\theta) = \log f(t|a;\theta) + \log f(a) \tag{2.2}$$

where $f(a)$ is the density for the ancillary.

It was shown in Fraser and Reid (1993) that in the same dimension case a crucial role is played by the so-called *tangent exponential model*, which is the exponential family model whose asymptotic expansion is closest to that of the true model.

The advantage is that highly accurate approximations available for the exponential family can be extended to general models.

In a full exponential family model, the canonical parameter can be obtained by differentiating the log likelihood function with respect to the minimal sufficient statistic. In the same way the canonical parameter of the tangent exponential model, is obtained by differentiating the log likelihood function of the conditional model with respect to the data variable for that model. However, as

$$\ell(\theta; y) = \ell(\theta; t|a) + \log f(a) \tag{2.2}$$

and the differentiation is for fixed a, the observed likelihood gradient from the conditional model can be computed from the gradient of the full model by differentiating in directions for which a is held fixed, that is, differentiating in directions tangent to the ancillary at the data point. Let $V = (v_1 \dots v_p)$ be p vectors tangent to the ancillary at the data, as given by say

$$V = \frac{\partial y(t,a)}{\partial t'}\Big|_{(t^0,a^0)} ;$$

also let

$$\frac{\partial \ell(\theta; y)}{\partial v_\alpha} = \frac{d}{dt}\ell(\theta; y + tv_\alpha)\Big|_{t=0}$$

be the directional derivative of the log likelihood function with respect to the data in the direction v_α. We then obtain the gradient of the conditional likelihood, denoted $\varphi'(\theta)$, by chain rule differentiation of (2.2):

$$\begin{aligned} \varphi'(\theta) = \ell'_{;t}(\theta; t^0\,|a^0) = \frac{\partial}{\partial t'}\ell(\theta, t|a^0)\Big|_{t=t^0} &= \ell_{;y'}(\theta; y^0)V = \sum_{i=1}^{n} \ell_{;y_i}(\theta; y^0)V_i \\ &= \frac{\partial}{\partial V}\ell(\theta, y)\big|_{y^0} = \ell_{;V}(\theta; y^0) \\ &= \left\{\frac{d}{dv_1}\ell(\theta; y^0), \dots, \frac{d}{dv_p}\ell(\theta; y^0)\right\} \end{aligned} \tag{2.3}$$

where V_i is the ith row of V. This gradient $\varphi'(\theta)$ gives a canonical reparametrization, which is the canonical parameter of the tangent exponential model at the data point y^0, an exponential model that gives the best fit to the conditional model at the data point (Fraser & Reid, 1995).

Dimension reduction from p to 1:

The reparametrization φ typically does not have the parameter of interest ψ as a linear component, so it is necessary to extract a linear surrogate for ψ from the new parameter φ. We obtain this by constructing a scalar parameter $\chi(\theta)$ that is a linear combination of the coordinates of $\varphi(\theta)$ and is tangent to $\psi(\theta)$ at $\hat{\theta}^0_\psi$:

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat{\theta}^0_\psi)}{|\psi_{\varphi'}(\hat{\theta}^0_\psi)|}\varphi(\theta) \ . \tag{2.4}$$

The gradient $\psi_\varphi(\theta)$ of $\psi(\theta)$ with respect to $\varphi(\theta)$ is calculated as

$$\psi_{\varphi'}(\theta) = \left\{\frac{\partial\psi(\theta)}{\partial\theta'}\right\}\left\{\frac{\partial\varphi(\theta)}{\partial\theta'}\right\}^{-1} = \psi_{\theta'}(\theta)\varphi^{-1}_{\theta'}(\theta) \ , \tag{2.5}$$

and is then evaluated at the constrained maximum likelihood value $\hat{\theta}^0_\psi$.

The quantity Q for use in (1.5) or (1.6) is then a standardized maximum likelihood departure for ψ calculated in the surrogate parametrization χ:

$$Q = \text{sgn}(\hat{\psi}^0 - \psi)\frac{|\chi(\hat{\theta}^0) - \chi(\hat{\theta}^0_\psi)|}{\hat{\sigma}_\chi(\varphi)} \tag{2.6}$$

where

$$\hat{\sigma}^2_\chi(\varphi) = \frac{|j_{(\lambda\lambda)}(\hat{\theta}^0_\psi)|}{|j_{(\theta\theta)}(\hat{\theta}^0)|}; \tag{2.7}$$

the argument φ in (2.7) is to indicate that the full and nuisance information determinants are recalibrated on the φ scale:

$$|j_{(\theta\theta)}(\hat{\theta}^0)| = |j_{\theta\theta}(\hat{\theta}^0)||\varphi_{\theta'}(\hat{\theta}^0)|^{-2} \ , \ |j_{(\lambda\lambda)}(\hat{\theta}^0_\psi)| = |j_{\lambda\lambda}(\hat{\theta}^0_\psi)||\varphi_{\lambda'}(\hat{\theta}^0_\psi)|^{-2}. \tag{2.8}$$

Note that Q involves an adjustment for the elimination of the nuisance parameter, which amounts to evaluating $j_{\lambda\lambda}$ at $\hat{\theta}_\psi$ rather than at $\hat{\theta}$ as in (1.3). This adjustment comes from using a Laplace approximation to evaluate the marginal distribution of a one-dimensional statistic that is ancillary for λ in the model with ψ fixed. This marginal distribution is exhibited on the curve $\hat{\theta}_\psi = \hat{\theta}^0_\psi$ of points that have the observed nuisance parameter maximum likelihood value. The exponential model approximation for this distribution has the canonical parameter (2.4) (Fraser & Reid, 1995).

2.2 Bayesian case

Tail probability formulas for the Bayesian context have been obtained by DiCiccio & Martin (1991). For the case with explicit nuisance parameterization they show that the marginal posterior survivor function for ψ can be approximated by (1.5) or (1.6) with r given by (1.4) and Q having the form of a standardized score statistic for ψ:

$$Q = \ell_\psi(\hat{\theta}_\psi) \left\{ \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}{|j_{\theta\theta}(\hat{\theta})|} \right\}^{\frac{1}{2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} \; , \tag{2.9}$$

where π is the prior density. The second factor appears as a reciprocal of that in (2.6) because the first factor is of score rather than maximum likelihood form. The third factor is a ratio of prior density values and can be called an adjustment factor; for some general results on such adjustments see Cheah, Fraser & Reid (1995). The effect of the prior is on the $O(n^{-1})$ term of the approximation. DiCiccio & Martin (1991) also present a more general formula that covers the case without nuisance parameterization but it has the disadvantage of treating the parameter coordinates asymmetrically and requiring the isolation of a coordinate θ_i such that $\partial\psi(\theta)/\partial\theta_i \neq 0$, with subsequent calculations specialized to this choice.

2.3 Comparison with r^*

It may be helpful to present a different expression for Q in (2.6) that emphasizes the connection to Barndorff-Nielsen's r^* approximation (Barndorff-Nielsen, 1986, 1991). The derivation given in the Appendix provides the following alternative expression to (2.6):

$$Q = \frac{|\ell_{;V'}(\hat{\theta}) - \ell_{;V'}(\hat{\theta}_\psi) \qquad \ell_{\lambda';V'}(\hat{\theta}_\psi)|}{|\ell_{\theta;V}(\hat{\theta})|} \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{\frac{1}{2}}, \tag{2.10}$$

the transposes here are needed to give correspondence to the use of row vectors by Barndorff-Nielsen. We note that this expression is independent of the choice of basis vectors V for the tangent space $\mathcal{L}(V)$ to the ancillary at the data point.

Barndorff-Nielsen's r^* approximation gives a p-value for ψ as $\Phi(r^*)$, where $r^* = r - r^{-1}\log(r/u)$, and is based on a quantity u given as (3.3) in Barndorff-Nielsen (1991) or as (6.108) in Barndorff-Nielsen & Cox (1994). The expression for u assumes that likelihood can written as $\ell(\theta; \hat{\theta}, a)$ and differentiation with respect to $\hat{\theta}$ is for fixed a; then using $\ell_{\theta;\hat{\theta}'}(\hat{\theta}) = j_{\theta\theta}(\hat{\theta})$ the expression for u can be rewritten as

$$u = \frac{|\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) \qquad \ell_{\lambda';\hat{\theta}}(\hat{\theta}_\psi)|}{|\ell_{\theta;\hat{\theta}'}(\hat{\theta})|} \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{\frac{1}{2}}. \tag{2.11}$$

By comparing (1.6) with $\Phi(r^*)$ we see that the formulas are identical if (2.10) and (2.11) are equal.

The equality of (2.10) and (2.11) based on the same ancillary a is then apparent by noting that differentiation with respect to $\hat{\theta}$ given a is in fact differentiation with respect to appropriate tangent vectors to the ancillary,

$$V = \frac{\partial y(\hat{\theta}, a)}{\partial \hat{\theta}'}$$

calculated at the data point.

In the special case that there is a minimal sufficient statistic t which is a one-to-one function of the maximum likelihood estimate $\hat{\theta}$, Q reduces to

$$Q = \frac{|\ell_{;t}(\hat{\theta}) - \ell_{;t}(\hat{\theta}_\psi) \qquad \ell_{\lambda';t}(\hat{\theta}_\psi)|}{|\ell_{\theta;t'}(\hat{\theta})|} \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{\frac{1}{2}} . \tag{2.12}$$

As Q is invariant under one-one transformations of t, we see directly the equivalence of Q and u.

In general the equality of (2.10) and (2.11) for given a shows that it is sufficient to compute the directional derivative of the log-likelihood function, and thus that it is not necessary to have an explicit expression for the exact or approximate ancillary statistic a. This makes Q easier to calculate and use in many problems. As well, the computation of the derivative with respect to $\hat{\theta}$ will usually require the use of sample space coordinates, and this is made explicit with the use of Q.

In the case of no nuisance parameters and thus a scalar parameter θ, (2.10) and (2.11) reduce respectively to

$$Q = \{\ell_{;V}(\hat{\theta}) - \ell_{;V}(\theta)\}\ell^{-1}_{\theta;V}(\hat{\theta})\{j(\hat{\theta})\}^{\frac{1}{2}}, \tag{2.13}$$

and to

$$u = \{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\theta)\}\{j(\hat{\theta})\}^{-\frac{1}{2}}$$

as noted in Reid (1996).

Another way to motivate the use of sample space differentiation, exemplified by (2.10) or (2.11), is to consider the derivation of (2.11) from the p^* approximation, outlined in Barndorff-Nielsen (1991) or the appendix of Pierce & Peters (1992). Consider the case that θ is a scalar. In order to use the density function

$$p(\hat{\theta}|a;\theta)$$

or an approximation to it for inference, we will typically want to compute

$$\int p(\hat{\theta}|a;\theta)\mathrm{d}\hat{\theta}.$$

The basic step in this derivation is a change of variable from $\hat{\theta}$ to r, so $\partial\ell/\partial\hat{\theta}$ arises as part of the Jacobian.

3. IMPLICIT NUISANCE PARAMETERIZATION

In this section we generalize the frequentist and Bayesian expressions for Q to the case where the scalar parameter of interest is given as $\psi(\theta)$ and there is no explicit nuisance parameterization. In the frequentist case the avoidance of nuisance parameterization in fact simplifies the expression and at the same time shows that the related background formulas are invariant with respect to nuisance parameterization. In the Bayesian case the general formula substantially simplifies the calculations for the case where $\psi(\theta)$ is nonlinear.

We do this by adapting the formulas of Section 2 to the implicit parametrization case. First, note that the constrained maximum likelihood value $\hat{\theta}_\psi$ is obtained by maximizing $\ell(\theta;y)+\alpha\{\psi(\theta)-\psi\}$ with respect to θ and α. Let

$$\tilde{\ell}(\theta)=\ell(\theta;y)+\hat{\alpha}\{\psi(\theta)-\psi\} \tag{3.1}$$

be the Lagrangian with the solution value $\hat{\alpha}$ substituted; it can be viewed as a tilted likelihood with maximum value at $\hat{\theta}_\psi$. A nominal information matrix $\tilde{\jmath}_{\theta\theta'}=\tilde{\jmath}_{\theta\theta'}(\hat{\theta}_\psi)=-\tilde{\ell}_{\theta\theta'}(\hat{\theta}_\psi)$ is calculated as the negative Hessian of $\tilde{\ell}(\theta)$ at $\hat{\theta}_\psi$.

We now define $\overline{\psi}(\theta)$ as an unnormalized version of the parameter χ defined in (2.4):

$$\overline{\psi}(\theta)=\psi_{\varphi'}(\hat{\theta}_\psi)\varphi(\theta)=\psi_{\theta'}(\hat{\theta}_\psi)\varphi_{\theta'}^{-1}(\hat{\theta}_\psi)\varphi(\theta) \tag{3.2}$$:

note carefully that the dependence on θ is only in the last factor $\varphi(\theta)$. The score parameter departure is just a derivative with respect to $\overline{\psi}$ or ψ at $\hat{\theta}_{\psi}$,

$$\left|\ell_{\overline{\psi}}(\hat{\theta}_{\psi})\right| = \frac{\left|\ell_{\theta}(\hat{\theta}_{\psi})\right|}{\left|\psi_{\theta}(\hat{\theta}_{\psi})\right|} , \tag{3.3}$$

and is well defined regardless of any choice of nuisance parametrization.

The variance estimate $\hat{\sigma}^2_{\psi}(\varphi)$ is determined in the φ parameterization and is of the same form as (1.3),

$$\hat{\sigma}^2_{\psi}(\varphi) = \{\psi_{\theta'}(\hat{\theta}_{\psi})\tilde{\jmath}^{\theta\theta}(\hat{\theta}_{\psi})\psi_{\theta}(\hat{\theta}_{\psi})\}\{\frac{|\tilde{\jmath}_{(\theta\theta)}|}{|\hat{\jmath}_{(\theta\theta)}|}\} , \tag{3.4}$$

where the information determinants have been recalibrated in the φ parameterization

$$|\hat{\jmath}_{(\theta\theta)}| = |\hat{\jmath}_{\theta\theta}||\varphi_{\theta}(\hat{\theta})|^{-2} , \quad |\tilde{\jmath}_{(\theta\theta)}| = |\tilde{\jmath}_{\theta\theta}(\hat{\theta}_{\psi})||\varphi_{\theta'}(\hat{\theta}_{\psi})|^{-2} . \tag{3.5}$$

It can happen that the two factors in braces in (3.4) are both negative, but they do have the same sign. Also some computational difficulties can arise if they are near zero; this latter condition can be avoided without altering the needed expression for Q by adding to (3.1) for computational purposes a term $-C^2\{\psi(\theta)-\psi\}^2$, a technique suggested in Hsu (1995). For the Bayesian case the estimated variance $\hat{\sigma}^2_{\psi}(\theta)$ is based on the parameterization chosen for integration here taken to be θ and thus the recalibration in (3.4) indicated by the parentheses in the subscripts is not needed.

The general formula gives the frequentist significance $p_f(\psi)$ and the Bayesian posterior probability $p_B(\psi)$ to third order, using (1.5) or (1.6) with the signed likelihood root r from (1.4) and the standardized parameter departure Q taken to be

$$q_f = \mathrm{sgn}(\hat{\psi}-\psi)\frac{|\overline{\psi}(\hat{\theta})-\overline{\psi}(\hat{\theta}_{\psi})|}{\hat{\sigma}_{\psi}(\varphi)} , \quad q_B = \mathrm{sgn}(\hat{\psi}-\psi)|\ell_{\overline{\psi}}(\hat{\theta}_{\psi})|\hat{\sigma}_{\psi}(\theta)\frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_{\psi})} \tag{3.6}$$

respectively for the frequentist and Bayesian cases, in parallel with (1.1) and (1.2). The scalar parameter $\overline{\psi}(\theta)$ is defined by (3.2), the score parameter departure $\ell_{\overline{\psi}}(\hat{\theta}_\psi)$ is defined by (3.3), and the variance estimate $\hat{\sigma}^2_\psi$ is given by (3.4) using in the frequentist case the exponential parameterization φ from (2.3) and using in the Bayesian case the parameterization θ chosen for the integration.

These formulae are established in the Appendix, using the results outlined in Section 2.

4. EXAMPLES

The general formula with (3.6) and (1.4) in (1.5) or (1.6) has been implemented in Maple V. As input the program takes the data vector or array y^0 and an expression in Maple language for the statistical model $f(y;\theta)$ and the parameter of interest $\psi(\theta)$. Component programs calculate the overall maximum likelihood value $\hat{\theta}^0$ and the constrained maximum likelihood value $\hat{\theta}^0_\psi$ for a chosen value $\psi(\theta) = \psi$, together with related informations $\hat{\jmath}$ and $\tilde{\jmath}$. The maximization programs are the computationally demanding ingredients and more robust versions would be appropriate.

For the frequentist case the program requires as input the pivotal quantity $z(y;\theta)$. The derivatives for (2.1) are determined using symbolic algebra, and substitution of the observed values then gives V. Further use of symbolic algebra then gives $\varphi(\theta)$, $\psi_\varphi(\theta)$, $\varphi_\theta(\theta)$ which with appropriate observed values give q_f in (3.6).

For the Bayesian case the prior $\pi(\theta)$ is required as input and components of q_B in (3.6) are obtained by symbolic algebra, with observed and restricted maximum likelihood estimates substituted as needed.

The preceding then gives the frequentist and Bayesian $p(\psi)$ for the chosen ψ value. Iterative application for different ψ values, then gives $p(\psi)$ as a function.

The Maple program is available at `www.utstat.utoronto.ca/reid/`. The present version requires additional input from the user in determining the grid of values over which to evaluate $p(\psi)$ and to monitor the computation of the restricted and unrestricted maximum likelihood estimates: this is the most difficult part of the calculation, but is also needed for most first order procedures.

Example 4.1: Normal coefficient of variation

Suppose we have a sample of size n from a normal distribution with mean μ and variance σ^2, and are interested in the coefficient of variation $\psi = \sigma/\mu$. A modified saddlepoint approximation is discussed in Vangel (1996) and illustrated there on the sample $(326, 302, 307, 299, 329)$. The transcript of the Maple session used to generate the significance functions using (1.5) or (1.6) with (1.4) and (3.6) is available on the above mentioned web site as an example. The ith pivotal quantity is taken to be $z_i = (y_i - \mu)/\sigma$ from which $V = (1, \hat{z}^0)$ where $\hat{z}^0$ is the observed standardized residual. The frequentist significance function is recorded in Figure 1, and the approximations (1.5) and (1.6) are very close in the plots. Figure 1 also shows the significance function calculated from the standard normal approximation for the signed square root of the log-likelihood ratio statistic, which in this case is not very accurate. The 95% confidence interval using (1.5) or (1.6) is $(0.0267, 0.1281)$: Vangel's approximation gives the interval $(0.0270, 0.1293)$. We compute the approximation from the sample y, but the same result is obtained if we reduce to the sufficient statistic first. However, a sufficiency reduction is not available if the error distribution is changed to, for example, Student or logistic, but the more general pivotal approach is easily implemented and just requires specifying the formula defining $f(y; \theta)$.

Figure 1 here

Example 4.2: Log-normal distribution

Suppose we have a sample from the normal distribution with mean μ and variance σ^2, and the parameter of interest is $\psi = \mu + \sigma^2/2$, which is the logarithm of the mean of the associated log-normal distribution. The pivotal is taken to be $z_i = (y_i - \mu)/\sigma$ as in Example (4.1). Figure 2(a) shows the significance function for ψ for the data of Lieblein and Zelen (1956) calculated using (1.5) and (1.6) and for comparison using the normal approximation for r. Figure 2(b) shows the significance function for $\log \sigma$. As in the previous example an initial reduction by sufficiency could be used, and would give the same result.

Figures 2(a) and 2(b) here

Example 4.3: Bayesian comparison of means

This Bayesian example is based on Hsu (1995, Section 4) and is a multivariate Behrens-Fisher problem. The data given in Hsu are the mean and variance of school expenditures per pupil, in five regions. Under a normal theory model, with different means and variances for each region and flat priors on the means and log-variances, the posterior marginal distribution for the means $\theta_1, \ldots, \theta_5$ is proportional to $\prod\{S_i^2 + n_i(\theta_i - \bar{y}_i)^2\}^{-n_i/2}$, where $\bar{y}_i$ is the sample mean and S_i^2 is the sum of squares within region i. Table 4.1 shows a set of nested posterior probability intervals for $\psi = \sum(\theta_i - \bar{\theta})^2$, using the normal approximation for r and using the general formula with (1.5). The marginal posterior density for ψ is plotted in Hsu (1995), and in line with the skewness evident in that density, the third order confidence intervals have a longer right tail than left tail; the posterior mode for ψ is 0.22. Hsu does not examine a frequentist p-value and correspondingly we do not; as the parameter ψ is curved in the location parameter we would expect the frequentist p-value to be different than the Bayesian p-value.

Table 4.1: Nested posterior probability intervals

90%	1st order	$(0.105, 0.355)$
	3rd order	$(0.110, 0.385)$
95%	1st order	$(0.090, 0.395)$
	3rd order	$(0.095, 0.435)$
98%	1st order	$(0.075, 0.455)$
	3rd order	$(0.080, 0.495)$
99%	1st order	$(0.065, 0.505)$
	3rd order	$(0.070, 0.565)$

Example 4.4: Inverse Gaussian distribution

Let $y_1, \ldots, y_n$ be a sample from the inverse Gaussian distribution with canonical parameters ψ and λ. The log-likelihood function is

$$\ell(\lambda, \psi) = \frac{n \log \psi}{2} + n(\psi\lambda)^{\frac{1}{2}} - \frac{\psi t_1}{2} - \frac{\lambda t_2}{2}$$

where $(t_1, t_2) = (\sum y_i^{-1}, \sum y_i)$ is the minimal sufficient statistic. No ancillary statistic is needed, as the maximum likelihood estimate is minimal sufficient. The maximum likelihood estimates satisfy

$$\left(\frac{\hat{\psi}}{\hat{\lambda}}\right)^{\frac{1}{2}} = \frac{t_2}{n}, \qquad \frac{1}{\hat{\psi}} + \left(\frac{\hat{\lambda}}{\hat{\psi}}\right)^{\frac{1}{2}} = \frac{t_1}{n}$$

and $\hat{\lambda}_\psi = \psi\hat{\lambda}/\hat{\psi}$; we also have $\ell_{;t}(\theta) = (-\psi/2, -\lambda/2)'$. Combining these with expressions for the Fisher information matrix and using (2.12) we obtain

$$Q = (\hat{\psi} - \psi)\left(\frac{n\psi}{2\hat{\psi}^3}\right)^{\frac{1}{2}},$$

which as expected is identical to the expression for u obtained by Barndorff-Nielsen (1990).

Example 4.5: Normal model with fixed coefficient of variation

Suppose $y_1, \ldots, y_n$ are independent and identically distributed from a normal distribution with mean θ and variance $b^2\theta^2$, where b is known. The log-likelihood function is

$$\ell(\theta) = -n\log\theta - \frac{nt_2}{2b^2\theta^2} + \frac{nt_1}{b^2\theta}$$

where $(t_1, t_2) = (\bar{y}, n^{-1}\sum y_i^2)$ is minimal sufficient. A vector of natural pivotal quantities z is $\{(y_1-\theta)/\theta, ..., (y_n-\theta)/\theta\}'$, which gives ancillary directions $V = (y_1/\hat{\theta}, \ldots, y_n/\hat{\theta})'$ and

$$\ell_{;V}(\theta) = -\frac{1}{b^2\theta^2}\sum(y_i-\theta)\frac{y_i}{\hat{\theta}} = \frac{n}{b^2\theta\hat{\theta}}(t_1 - \frac{t_2}{\theta})$$

so that

$$Q = \left\{t_1(\frac{1}{\hat{\theta}} - \frac{1}{\theta}) - t_2(\frac{1}{\hat{\theta}^2} - \frac{1}{\theta^2})\right\}\left(\frac{n^{\frac{1}{2}}\hat{\theta}}{b}\right)(t_2 + b^2\hat{\theta}^2)^{-\frac{1}{2}}$$

The same result is obtained working directly from the minimal sufficient statistic using the pivotal $\{(t_1-\theta)/\theta, t_2/\theta^2\}$.

There is an exact ancillary statistic for this model (Hinkley, 1977), which is $t_2^{\frac{1}{2}}/t_1$, and the maximum likelihood estimate can be given explicitly as the root of a quadratic equation, so it is not difficult to calculate u in this case either. The exact conditional density for t_2 given the ancillary is provided in Hinkley (1977), and a recursion can be developed for the exact cumulative distribution function for $\hat{\theta}$. Table 5.1 below compares approximate values from (1.6), using Q in (3.6) or (2.10) for selected values of b^2, with the exact values based on the ancillary used by Hinkley. Scale invariance shows easily that the tangent direction V leading to Q is tangent to the Hinkley ancillary and thus that the third order approximation is directly an approximation to the exact value based on that ancillary. Table 4.2 also shows the approximation to the p-values using Skovgaard's (1996) statistic $\tilde{U}$. This approximation has relative accuracy $O(n^{-1})$ in a large deviation neighbourhood.

Table 4.2. Exact and approximate p-values: normal coefficient of variation.

$b = 1$

θ	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$\Phi(r)$	0.99601	0.84653	0.50145	0.24136	0.11021	0.05126	0.02493	0.01277
Skovgaard	0.99751	0.88368	0.57429	0.30528	0.15253	0.07675	0.03998	0.02174
(1.6)	0.99757	0.88546	0.57840	0.30938	0.15553	0.07871	0.04122	0.02253
exact	0.99760	0.88627	0.58016	0.31102	0.15665	0.07940	0.04163	0.02277

$b = 2$

θ	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2
$\Phi(r)$	0.99232	0.90157	0.69320	0.46892	0.29748	0.18475	0.11499	0.04673
Skovgaard	0.99517	0.92855	0.75417	0.54426	0.36792	0.24230	0.15904	0.07077
(1.6)	0.99535	0.93058	0.75935	0.55131	0.37505	0.24852	0.16407	0.07377
exact	0.99534	0.93105	0.76046	0.55270	0.37635	0.24956	0.16484	0.07416

$b = 1/2$

θ	0.8	1	1.2	1.4	1.6	1.8	2.0	2.2
$\Phi(r)$	0.99189	0.80553	0.42601	0.17110	0.06257	0.02294	0.00880	0.00359
Skovgaard	0.99452	0.84502	0.49142	0.21907	0.08836	0.03540	0.01470	0.00643
(1.6)	0.99452	0.84498	0.49134	0.21901	0.08832	0.03538	0.01469	0.00643
exact	0.99456	0.84582	0.49291	0.22028	0.08907	0.03576	0.01488	0.00652

Example 4.6: A (2,1) exponential family

Let y_1 and $y_2 - 1$ be independent and exponentially distributed with parameters χ and ψ so that the joint probability density function of y_1 and y_2 is

$$f(y_1, y_2) = \chi\psi e^{\psi} e^{-\chi y_1 - \psi y_2}. \tag{4.1}$$

Barndorff-Nielsen & Chamberlin (1991) considered third order inference for the submodel of (4.1) specified by $\chi = \chi(\psi) = \psi^{-1}e^{-\psi}$. As no exact ancillary statistic exists for this model, they used the affine ancillary statistic (Barndorff-Nielsen, 1986) to calculate u in (2.11). For the submodel, the pivotal quantity $z = \{\chi(\psi)y_1, \psi(y_2 - 1)\}$ gives ancillary directions $V = \{(1+\hat{\psi}^{-1})y_1, -\hat{\psi}^{-1}(y_2-1)\}$. The likelihood gradient is $\ell_{;V}(\psi) = -\chi(\psi)(1+\hat{\psi}^{-1})y_1 + \psi\hat{\psi}^{-1}(y_2-1)$ and we obtain $\ell_{\psi;V}(\hat{\psi}) = \chi(\hat{\psi})(1+\hat{\psi}^{-1})^2 y_1 + \hat{\psi}^{-1}(y_2-1)$. Combining this with the expression for the observed information gives

$$Q = \left[\left\{\frac{\chi(\psi)}{\chi(\hat{\psi})} - \frac{\psi}{\hat{\psi}}\right\}y_2 + \frac{\psi}{\hat{\psi}}\right]\left(y_2 + \frac{2y_2}{\hat{\psi}} - \frac{1}{\hat{\psi}}\right)^{-1}\left\{\frac{y_2(2/\hat{\psi}+2+\hat{\psi})}{1+\hat{\psi}}\right\}^{\frac{1}{2}}. \qquad (4.2)$$

Table 4.3 gives the results from 20,000 simulations of a sample of size 1 from the curved exponential model with $\psi = 1$. We compare at selected points the empirical distribution function F_r for r, the empirical distribution function F_u^* for r^* using u in (2.11) from Barndorff-Nielsen & Chamberlin (1991), the empirical distribution function F_Q^* for r^* using Q in (4.2), and the target distribution which is $N(0,1)$.

Table 4.3. The empirical cdf of r, r^* using u, and r^* using Q, compared to the $N(0,1)$ distribution

x	-3.0	-1.96	-0.67	0.0	0.67	1.96	3.0
F_r	0.0009	0.0208	0.311	0.593	0.824	0.9874	0.9993
F_u^*	0.0012	0.0229	0.236	0.485	0.742	0.9759	0.9987
F_Q^*	0.0012	0.0221	0.227	0.480	0.740	0.9759	0.9987
$\Phi(x)$	0.0013	0.025	0.251	0.500	0.749	0.9750	0.9987

Example 4.7: Equality of gamma means

We consider a sample of size n_1 from a Gamma distribution with shape β_1 and mean μ, and an independent sample of size n_2 from a Gamma distribution with shape β_2 and mean μ. The log-likelihood function is

$$\begin{aligned}\ell(\beta_1,\beta_2,\mu) = {} & -n_1\log\Gamma(\beta_1) - n_2\log\Gamma(\beta_2) - (n_1\beta_1 + n_2\beta_2)\log\mu \\ & + n_1\beta_1\log\beta_1 + n_2\beta_2\log\beta_2 + \beta_1\sum\log y_{1i} + \beta_2\sum\log y_{2i} \\ & - (\beta_1/\mu)\sum y_{1i} - (\beta_2/\mu)\sum y_{2i}\end{aligned}$$

which is a (4,3) curved exponential model. Although there is a reduction to a four dimensional sufficient statistic, it is easier to work with the original observations $(y_{11},\ldots,y_{1n_1},y_{21},\ldots,y_{2n_2})$. A natural pivotal vector $z=(z_{11},\ldots,z_{1n_1},z_{21},\ldots,z_{2n_2})$ has entries $z_{ji}=F(y_{ji};\beta_j,\mu)$, where $F(y;\beta,\mu)=F(\beta y/\mu;\beta)$ and $F(z;\beta)$ is the Gamma distribution function with shape β and mean 1.

The matrix V from (2.1) needed to define φ in (2.3) is $(n_1+n_2)\times 3$ with rows designed $(1,i)$ and $(2,i)$ given as,

$$V_{1i}=\{c(\hat\beta_1,\hat\mu;y_{1i}),0,y_{1i}/\hat\mu\}$$

$$V_{2i}=\{0,c(\hat\beta_2,\hat\mu,y_{2i}),y_{2i}/\hat\mu\}$$

where $c(\beta,\mu;y)=-F_{\boldsymbol{\beta}}(y;\beta,\mu)/F_{\boldsymbol{y}}(y;\beta,\mu)$. We can then compute $\varphi(\beta_1,\beta_2,\mu)$ which is the transpose of $\ell_{;V}$

$$\begin{pmatrix} \beta_1\sum y_{1i}^{-1}c(\hat\beta_1,\hat\mu,y_{1i})-(\beta_1/\mu)\sum c(\hat\beta_1,\hat\mu,y_{1i}) \\ \beta_2\sum y_{2i}^{-1}c(\hat\beta_2,\hat\mu,y_{2i})-(\beta_2/\mu)\sum c(\hat\beta_2,\hat\mu,y_{2i}) \\ \beta_1(n_1/\hat\mu)-(\beta_1/\mu)\sum y_{1i}+\beta_2(n_2/\hat\mu)-(\beta_2/\mu)\sum y_{2i} \end{pmatrix},$$

and is of course linear in the canonical parameters β_1, β_2, β_1/μ, β_2/μ. These are readily combined into expressions (1.5) and (3.6).

ACKNOWLEDGMENTS

The authors wish to express deep appreciation for the insightful and helpful comments of an Associate Editor and a referee which significantly modified and improved this paper.

APPENDIX: Technical details

Derivation of (3.6): frequentist case

For the frequentist case we derive the general formula in Section 3 from the special formula in Section 2.1 where an explicit nuisance parameterization is needed.

For this we have the interest parameter $\psi(\theta)$, and we let λ be some underlying nuisance parametrization so that temporarily we can write $\theta' = (\lambda', \psi)$. First we note that the likelihood ratio r does not involve the nuisance parameterization, and is the same for both the special and the general formula. Second, we show that Q in (2.6) gives the q_f in (3.6). The estimated variance for $\chi(\theta)$ is given by (2.7) and can be modified using the matrix algebra result implicit in (1.3):

$$\begin{aligned}\hat{\sigma}^2_{\chi}(\varphi) &= \frac{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|}{|j_{(\theta\theta)}(\hat{\theta})|} = \frac{\tilde{\jmath}^{(\psi\psi)}(\hat{\theta}_\psi)|\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)|}{|j_{(\theta\theta)}(\hat{\theta})|} \\ &= \chi_{\theta'}(\hat{\theta}_\psi)\tilde{\jmath}^{\theta\theta}\chi_\theta(\hat{\theta}_\psi)\frac{|\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)|}{|j_{(\theta\theta)}(\hat{\theta})|} \ .\end{aligned}$$

The estimated variance $\sigma^2_{\psi}(\varphi)$ for the basic parameter $\psi(\theta)$ or $\bar{\psi}(\theta)$ is then just $|\psi_{\varphi'}(\hat{\theta}_\psi)|^2$ times the above estimated variance for $\chi(\varphi)$ and is thus given by (3.4). Correspondingly the parameterizations $\chi(\theta)$ and $\overline{\psi}(\theta)$ differ by the scale factor $|\psi_{\varphi'}(\hat{\theta}_\psi)|$. When these re-expressions are substituted into (2.6), the scale factors cancel giving q_f in (3.6). Also from this it follows that the formulas are parametrization invariant.

Derivation of (3.6): Bayesian case

For the Bayesian case we again need to show only that the special case Q in (2.9) for an explicit nuisance parameterization (λ', ψ) can be transformed to give the general Q in (3.6). To do this we first need to generalize the assumptions and the derivation of the Bayesian Q in (2.9) and for this follow methods in Fraser and Reid (1995).

The probability $p_B(\psi)$ and the formulas for Q and r need only the surface $\psi(\theta) = \psi$, together with a positive direction from that surface, and thus do not depend otherwise on the function $\psi(\theta)$. This permits simplifications. Let T be the tangent plane to $\psi(\theta) = \psi$ at $\hat{\theta}^0_\psi$ and P be the corresponding perpendicular at that point. These lead to alternate coordinates on the parameter space.

Consider a Euclidean coordinate or distance α along P taking for convenience the value ψ at $\hat{\theta}^0_{\boldsymbol{\psi}}$. Then more generally for θ on a plane parallel to T, let $\alpha(\theta)$ be the just-described Euclidean coordinate at the intersection of the plane and P. We then have that $\hat{\theta}^0_{\boldsymbol{\psi}}$ is the same whether calculated from $\psi(\theta) = \psi$ or $\alpha(\theta) = \psi$.

Now consider a transformation on the parameter space. Let $C = \{\hat{\theta}_{\boldsymbol{\alpha}}\}$ be the curve of constrained maximum likelihood values given $\alpha(\theta) = \alpha$. Then translate each plane $\alpha(\theta) = \alpha$ to carry $\hat{\theta}_{\boldsymbol{\alpha}}$ into the α point on P; the new curve $C = \{\hat{\theta}_{\boldsymbol{\alpha}}\}$ is then the perpendicular P. The full transformation is volume preserving and also does not alter the Bayesian Q in (3.6).

Next we define a new or canonical $\psi(\theta)$ that takes the value ψ on the initial surface $\psi(\theta) = \psi$ and increments as does $\alpha(\theta)$ in the direction P; the contours of the new $\psi(\theta)$ are then translates of $\psi(\theta) = \psi$ in the direction P. To complete the general coordinates, we now let λ give orthogonal Euclidean coordinates perpendicular to P.

In the new coordinates we then apply the methods in DiCiccio & Martin (1991) and Fraser & Reid (1995) and obtain to third order the relative density

$$\exp\{\ell(\hat{\theta}_{\boldsymbol{\psi}})\}|j_{\boldsymbol{\lambda\lambda}}(\hat{\theta}_{\boldsymbol{\psi}})|^{-\frac{1}{2}}\pi(\hat{\theta}_{\boldsymbol{\psi}})$$

for the marginal distribution of $\psi(\theta)$ on the line P; the second factor comes from the nuisance parameter integration. In the modified coordinates we have that this the same as if the initial θ' were given as (λ', ψ); accordingly we can then directly apply the earlier (2.9).

The gradient $\ell_{\boldsymbol{\psi}}(\hat{\theta}_{\boldsymbol{\psi}})$ of the likelihood at $\hat{\theta}_{\boldsymbol{\psi}}$ can be obtained at that point by partial differentiation given any nuisance parameterization and can be expressed as $\ell_{\overline{\boldsymbol{\psi}}}(\hat{\theta}_{\boldsymbol{\psi}})$ in (3.3) without recourse to a nuisance parameterization. The information adjustment in (2.9) can be reexpressed

$$\begin{aligned}\frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}{|j_{\theta\theta}(\hat{\theta})|} &= \frac{\tilde{j}^{\psi\psi}(\hat{\theta}_\psi)|\tilde{j}_{\theta\theta}(\hat{\theta}_\psi)|}{|j_{\theta\theta}(\hat{\theta})|} \\ &= \psi_{\theta'}(\hat{\theta}_\psi)\tilde{j}^{\theta\theta}(\hat{\theta}_\psi)\psi_\theta(\hat{\theta}_\psi)\frac{|\tilde{j}_{\theta\theta}|}{|\hat{j}_{\theta\theta}|} \\ &= \hat{\sigma}^2_\psi(\theta) \ ,\end{aligned}$$

giving the estimated variance (3.4) but without the re-calibration indicated by the parentheses in the subscripts.

Derivation of (2.10)

The lead determinant in the expression (2.10) for Q can be rewritten in terms of the φ notation introduced by (2.3):

$$|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \qquad \varphi_{\lambda'}(\hat{\theta}_\psi)| \ .$$

The determinant records the volume of the parallelpiped spanned by the p column vectors in the array. For this the first vector can be orthogonalized to the remaining $p-1$ vectors; the value of the determinant is then the length of the orthogonalized first vector times the determinant $|\varphi_{\lambda'}(\hat{\theta}_\psi)| = |\varphi'_\lambda(\hat{\theta}_\psi)\varphi_{\lambda'}(\hat{\theta}_\psi)|^{\frac{1}{2}}$ implicit in (2.8). As the vectors $\varphi_{\lambda'}(\hat{\theta}_\psi)$ are tangent to $\psi(\theta) = \psi$ at $\hat{\theta}_\psi$ in the φ-space, it follows that the length of the orthogonalized first vector is just $|\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)|$ in (2.6) based on (2.4). We then note that $|\ell_{\theta;V'}(\hat{\theta})| = |\varphi_{\theta'}(\hat{\theta})|$ gives the recalibration taking $|j_{\theta\theta}(\hat{\theta})|^{\frac{1}{2}}$ to $|j_{(\theta\theta)}(\hat{\theta})|^{\frac{1}{2}}$, and the $|\varphi_{\lambda'}(\hat{\theta}_\psi)|$ from the lead determinant gives the recalibration taking $|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{\frac{1}{2}}$ to $|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{\frac{1}{2}}$. This gives the equivalence of (2.6) and (2.10).

REFERENCES

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-365.

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307-322.

Barndorff-Nielsen, O.E. (1990). Approximate interval probabilities. *J.R. Statist. Soc. B* **52**, 485-496.

Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557-563.

Barndorff-Nielsen, O.E. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *J.R. Statist. Soc. B* **56**, 125-140.

Barndorff-Nielsen, O.E. & Chamberlin, S. (1991). An ancillary invariant modification of the signed log likelihood ratio. *Scand. J. Statist.* **18**, 341-352.

Barndorff-Nielsen, O.E. & Chamberlin, S. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81**, 485-494.

Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics.* London, Chapman and Hall.

Cakmak, S., McDunnough, P., Reid, N. & Yuan, X. (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Plan. Inf.* **66**, 211-222.

Cheah, P.K., Fraser, D.A.S. & Reid, N. (1995). Adjustment to likelihood and densities; calculating significance. *J. Statist. Res.* **29**, 1-13.

DiCiccio, T., Field, C. & Fraser, D.A.S. (1990). Marginal tail probabilities and inference for real parameters, *Biometrika* **77**, 65-76.

DiCiccio, T. & Martin, M.A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika* **78**, 891-902.

DiCiccio, T. & Martin, M.A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. R. Statist. Soc.* B **55**, 305-318.

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

Fraser, D.A.S. & Reid, N. (1993). Simple asymptotic connections between density and cumulant functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67-82.

Fraser, D.A.S. & Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **7**, 33-53.

Fraser, D.A.S. & Reid, N. (1998). Ancillary information for statistical inference. *Proceedings of a Symposium on Empirical Bayes and Likelihood Inference.* Springer-Verlag, New York, to appear.

Hinkley, D.V. (1977). Conditional inference about a normal mean with known coefficient of variation. *Biometrika* **64**, 105-108.

Hsu, J. (1995). Generalized Laplacian approximations in Bayesian inference. *Canad. J. Statist.* **23**, 399-410.

Lieblein, J. & Zelen, M. (1956). Statistical investigation of the fatigue life of deep groove ball bearings. *J. Research, National Bureau of Standards* **57**, 273-316.

Lugannani, R. & Rice, S. (1980). Saddlepoint approximation for the distribution function of the sum of independent variables. *Adv. Appl. Prob.* **12**, 475-490.

Pierce, D.A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *J.R. Statist. Soc. B* **54**, 701-737.

Reid, N. (1996). Likelihood and higher order approximations to tail areas: A review and annotated bibliography. *Canad. J. Statist.* **24**, 141-166.

Skovgaard, I.M. (1986). Successive improvements of the order of ancillarity. *Biometrika* **73**, 516-519.

Skovgaard, I.M. (1996). An explicit large-deviation approximation to one parameter tests. *Bernoulli* **2**, 145-165.

Vangel, M.G. (1996). Confidence intervals for a normal coefficient of variation.

Amer. Statist. **15**, 21-26.

Figure 1: Significance probability for normal coefficient of variation (Example 4.1). Solid line: normal approximation to r; Dotted line: approximation (1.5); Dashed line: approximation (1.6).

Figure 2(a): Significance probability for log-normal mean (Example 4.2). Solid line: normal approximation to r; Dotted and dashed lines: approximations (1.5) and (1.6).

Figure 2(b): Significance probability for log variance (Example 4.2). Solid line: normal approximation to r; Dotted and dashed lines: approximations (1.5) and (1.6).