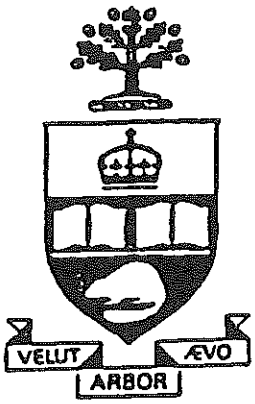

A simple general formula for tail probabilities for frequentist and Bayesian inference

by

D.A.S. Fraser
Department of Statistics
University of Toronto

and

N. Reid
Department of Statistics
University of Toronto

and

J. Wu
Department of Statistics
University of Toronto

Technical Report No. 9612, December 3, (1996)

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

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

D.A.S. Fraser, N. Reid and J. Wu
Department of Statistics
University of Toronto
Toronto, Ontario M5S 3G3

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. The relationship to Barndorff-Nielsen's r^* approximation is discussed.

Some Keywords: Bayesian tail probabilities; Frequentist tail probabilities; Scalar parameters; Third order inference

departure and requires some specialized notation which is defined in Section 3.

The general formula is easily implemented and requires the specification of the statistical model $f(y;\theta)$, the data y^0, the interest parameter $\psi(\theta)$, and a pivotal quantity of dimension equal that of the variable y. In the Bayesian case a model $L(\theta - y)$, with data $y = 0$ is input together with the interest parameter $\psi(\theta)$ and prior $\pi(\theta)$.

The formulas of Fraser & Reid (1995) and DiCiccio & Martin (1991) are described in Section 2, and these are used to verify the general formula in Section 3. Some examples are presented in Section 4. A comparison of the Fraser & Reid (1995) formula to that of Barndorff-Nielsen (1991) is given in Section 5.

2. WITH EXPLICIT NUISANCE PARAMETRIZATION

2.1 Frequentist case

The verification of the general formula in the frequentist case is based on a formula in Fraser & Reid (1995). We assume that the model $f(y;\theta)$ has asymptotic properties so that the log-likelihood function $\ell(\theta;y)$ is $O(n)$ in θ, that the maximum likelihood estimate $\hat{\theta}$ converges at rate $1/\sqrt{n}$ to θ, and so on. We assume here that we have an explicit nuisance parametrization $\theta' = (\lambda', \psi)$ where ψ is the scalar interest parameter and that the dimension of y is greater than or equal to that of θ.

As primary input the formula uses the signed likelihood root

$$R = \text{sgn}(\hat{\psi}^0 - \psi) \cdot [2\{\ell(\hat{\theta}^0; y^0) - \ell(\hat{\theta}^0_\psi; y^0)\}]^{\frac{1}{2}} \tag{2.1}$$

where y^0 is the observed value of y; this is the same as in (1.3), but emphasizes that the significance probability is computed for a fixed data value.

To compute the complementing quantity Q, we need a vector q of n pivotal quantities q_i where $q_i = q(y_i;\theta)$ depends on y_i, has a fixed distribution, and indicates how each coordinate is measuring the parameter θ. Some examples of this are provided in the

examples in Section 4. We then construct an array $V = (v_1 \cdots v_p)$ of p linearly independent vectors in R^n using q:

$$V = \frac{dy}{d\theta}\Big|_{(y^0,\hat{\theta}^0)} = (\frac{\partial q}{\partial y})^{-1}(\frac{-\partial q}{\partial \theta})\Big|_{(y^0,\hat{\theta}^0)} \tag{2.2}$$

where the first expression is calculated for fixed q, and y^0 is the observed data point with maximum likelihood estimate $\hat{\theta}^0$. As discussed in detail in Fraser & Reid (1995), the vectors $(v_1, \ldots, v_p)$ are tangent to a location-based ancillary, and also tangent to a modification of this location-based ancillary that is ancillary to $O(n^{-1})$. It is shown there, for coordinates each with a scalar parameter and more generally in work as yet unpublished, that this is sufficient for the approximation to the significance probability to have relative error $O(n^{-3/2})$, and in particular that it is not necessary to actually compute or know the ancillary statistic. It is this simplification that allows the formula to be fairly easily implemented.

The array V is used to obtain a nominal reparametrization

$$\begin{aligned} \varphi'(\theta) = \ell_{;V}(\theta; y^0) &= \frac{d}{dV}\ell(\theta; y)\Big|_{y^0} = \left(\frac{d}{dv_1}\ell(\theta; y), \ldots, \frac{d}{dv_p}\ell(\theta; y)\right)\Big|_{y^0} \\ &= \sum \ell_{;y_i}(\theta; y)\Big|_{y^0} V_i, \end{aligned} \tag{2.3}$$

where V_i is the ith row of V. The reparametrization φ records the gradient of the likelihood function in the ancillary directions V. This assumes a continuous statistical model to permit the needed sample space differentiation. Note that this reparametrization is specific to the observed data point y^0, which we emphasize in (2.1) and in the following formulas.

The next step is to construct from φ a scalar parameter $\chi(\theta)$ that replaces $\psi(\theta)$:

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

This parameter is a linear combination of the coordinates of $\varphi(\theta)$ using the Jacobian of $\psi(\theta)$ with respect to $\varphi(\theta)$ at the constrained maximum likelihood value $\hat{\theta}^0_\psi$. In (2.4)

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

The secondary input Q is the standardized maximum likelihood departure

$$Q = \text{sgn}(\hat{\psi}^0 - \psi)|\chi(\hat{\theta}^0) - \chi(\hat{\theta}^0_\psi)| \left\{ \frac{|j_{(\theta\theta)}(\hat{\theta}^0)|}{|j_{(\lambda\lambda)}(\hat{\theta}^0_\psi)|} \right\}^{\frac{1}{2}} \tag{2.6}$$

where $|j_{(\theta\theta)}(\hat{\theta}^0)|$ and $|j_{(\lambda\lambda)}(\hat{\theta}^0_\psi)|$ are full and nuisance information determinants, recalibrated on the φ scale:

$$|j_{(\theta\theta)}(\hat{\theta}^0)| = |j_{\theta\theta}(\hat{\theta}^0)||\varphi_\theta(\hat{\theta}^0)|^{-2} \ , \ |j_{(\lambda\lambda)}(\hat{\theta}^0_\psi)| = |j_{\lambda\lambda}(\hat{\theta}^0_\psi)||\varphi_{\lambda'}(\hat{\theta}^0_\psi)|^{-2}; \tag{2.7}$$

we use the notation $|X| = |X'X|^{\frac{1}{2}}$ for the determinant of a $p \times (p-1)$ matrix X.

2.2 Bayesian case

The verification of the general model for the Bayesian context is based on a tail probability formula in DiCiccio & Martin (1991). Although DiCiccio & Martin (1991) give a version that does not require explicit nuisance parametrization, we assume here that $\theta' = (\lambda', \psi)$ which is more convenient for the present generalization. DiCiccio & Martin (1991) show that the marginal posterior survivor function or right tail distribution function for ψ can be approximated by (1.1) or (1.2) with R given by (1.3) and Q having the form of a standardized score statistic for ψ:

$$Q = \ell_\psi(\hat{\theta}_\psi) \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{-\frac{1}{2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} \ ; \tag{2.8}$$

where π is the prior density. The second factor appears as a reciprocal of that in (2.6) in part 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 et al (1995). The disadvantage of a more general formula in DiCiccio & Martin (1991) is that it treats the parameter coordinates asymmetrically and requires the isolation of a coordinate θ_i such that $\partial\psi(\theta)/\partial\theta_i \neq 0$, with subsequent calculations specialized to this choice.

3. GENERAL TAIL PROBABILITY FORMULA

3.1 The formula

In this Section we generalize the formulas from Section 2 to the case where the scalar parameter of interest is given as $\psi(\theta)$ without explicit nuisance parameterization. And for notational convenience we suppress the dependence of the formulas for R and Q on the observed data point y^0 or $\hat{\theta}^0$.

The tail probability $p(\psi)$ for assessing the value ψ is given by (1.1) or (1.2) with signed likelihood root R given by (1.3) and with standardized maximum likelihood departure given by

$$Q = \mathrm{sgn}(\hat{\psi} - \psi)|\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)| \left\{ \frac{|\hat{\jmath}_{(\theta\theta)}(\hat{\theta})|}{\hat{\sigma}^2(\chi)|\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)|} \right\}^{\frac{1}{2}} \frac{a(\hat{\theta})}{a(\hat{\theta}_\psi)} \,. \tag{3.1}$$

We now record details for the ingredients in the expression for Q. The overall maximum likelihood value $\hat{\theta}$ is obtained by maximizing the log likelihood $\ell(\theta; y) = \log f(y; \theta)$. The ancillary directions V are obtained from componentwise pivotal quantities as at (2.2). In models such as canonical exponential families, where a reduction by sufficiency is available, it suffices to use $V = I$; and for transformation models it suffices to use the standard error pivotal although the usual conditional procedure would give the same result. The new parameterization $\varphi(\theta)$ is given by (2.3). The scalar parameter $\chi(\theta)$ is linear in the φ parameterization,

$$\chi(\theta) = \psi_{\varphi'}(\hat{\theta}_\psi)\varphi(\theta) \tag{3.2}$$

where $\psi_{\varphi'}(\theta)$ is given by (2.5): this omits the normalization used in (2.4).

The constrained maximum likelihood value $(\hat{\theta}_\psi, \hat{\alpha})$ is obtained by maximizing $\ell(\theta; y) + \alpha\{\psi(\theta) - \psi\}$. Let

$$\tilde{\ell}(\theta) = \ell(\theta; y) + \hat{\alpha}\{\psi(\theta) - \psi\} \tag{3.3}$$

be the Lagrangian in the calculations; it can be viewed as a tilted likelihood with maximum at $\hat{\theta}_\psi$. At $\hat{\theta}$ we use the information determinant (2.7). At $\hat{\theta}_\psi$ we use a nominal information matrix

$$\tilde{\jmath}_{\theta\theta}(\hat{\theta}_\psi) = -\tilde{\ell}_{\theta\theta}(\hat{\theta}_\psi) \tag{3.4}$$

and its inverse $\tilde{j}^{\theta\theta}(\hat{\theta}_\psi)$. This gives a recalibrated information determinant

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

and an estimated nominal variance

$$\hat{\sigma}^2(\chi) = \psi_{\theta'}(\hat{\theta}_\psi)\tilde{j}^{\theta\theta}(\hat{\theta}_\psi)\psi_\theta(\hat{\theta}_\psi) \tag{3.6}$$

for the parameter χ. It can happen that expressions (3.5) and (3.6) are both negative, but they always have the same sign. More computational difficulty arises if they are near zero; this latter condition can be avoided without altering the needed expression for Q by adding to (3.3) for computational purposes a term $-C^2\{\psi(\theta) - \psi\}$, a technique suggested in Hsu (1995). The function $a(\theta)$ is the prior density $\pi(\theta)$ in the Bayesian case. In our frequentist examples $a(\theta)$ to be 1, but in other frequentist contexts it could be useful to allow incorporation of an adjustment factor of this form.

2.2 The frequentist verification

In the frequentist setting, we take $a(\theta) = 1$, and formula (3.1) is derived from (2.6) as follows. Let $(\bar{\varphi}_1, \ldots, \bar{\varphi}_p)$ be a rotation of $(\varphi_1, \ldots, \varphi_p)$ so that the curves $\bar{\varphi} = constant$ and $\psi = constant$ are tangent at $\hat{\theta}_\psi$. We then have that

$$\bar{\theta}'(\theta) = \{\bar{\varphi}_1(\theta), \ldots, \bar{\varphi}_{p-1}(\theta), \psi(\theta)\} = \{\lambda'(\theta), \psi(\theta)\}$$

provides a reparametrization appropriate to the tested value ψ with an explicit nuisance parameter λ. Let $\tilde{j}_{\bar{\theta}\bar{\theta}}(\bar{\theta})$ be the corresponding nominal information matrix obtained from $\tilde{j}_{\theta\theta}(\theta)$ but recalibrated on the ϕ scale. Then $|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|$ in (2.7) is equal to $|\tilde{j}_{\lambda\lambda}(\hat{\theta}_\psi)|$ since $\bar{\varphi}$ is a rotation of φ, and $\bar{\varphi}_\lambda(\hat{\theta}_\psi)$ is just the identity. From a standard identity for determinants of partitioned matrices we then have

$$|\tilde{j}_{\lambda\lambda}(\hat{\theta}_\psi)| = |\tilde{j}_{\bar{\theta}\bar{\theta}}(\hat{\theta}_\psi)||\tilde{j}^{\psi\psi}(\hat{\theta}_\psi)| = |\tilde{j}_{\bar{\theta}\bar{\theta}}(\hat{\theta}_\psi)|\hat{\sigma}^2(\chi),$$

using (3.6). We thus obtain (3.1) from (2.6) by noting that

$$\begin{aligned}
|\tilde{\jmath}_{\bar{\theta}\bar{\theta}}(\hat{\theta}_\psi)| &= |\tilde{\jmath}_{\theta\theta}(\hat{\theta}_\psi)||\theta_{\bar{\theta}'}(\hat{\theta}_\psi)|^2 \\
&= |\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)||\varphi_{\theta'}(\hat{\theta}_\psi)|^2|\theta_{\bar{\theta}'}(\hat{\theta}_\psi)|^2 \\
&= |\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)||\varphi_{\bar{\theta}'}(\hat{\theta}_\psi)|^2 \\
&= |\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)||\bar{\theta}_{\varphi'}(\hat{\theta}_\psi)|^{-2} \\
&= |\tilde{\jmath}_{(\theta\theta)}(\hat{\theta}_\psi)||\psi_{\varphi'}(\hat{\theta}_\psi)|^{-2},
\end{aligned}$$

the last line following because $\bar{\varphi}$ is a rotation of φ. This provides the missing normalization between the definitions (2.4) and (3.3).

3.3 The Bayesian verification

To apply the formula in the Bayesian case, we use the likelihood function $L(\theta)$, and consider the model for computation as $f(t;\theta) = L(\theta - t) = \exp\{\ell(\theta - t)\}$ with data value $t^0 = 0$, where θ and t have the same dimension p. We verify (3.1) from (2.8).

The values of R and Q are unaffected by a rotation of coordinates for θ and t, and accordingly we rotate the coordinates so that the curve $\theta_p = constant$ is tangent with the curve $\psi(\theta) = constant$ at $\hat{\theta}_\psi$. The new parameterization $\varphi(\theta)$ is obtained by differentiating $\ell(\theta - t)$ with respect to t, or equivalently, with respect to $-\theta$:

$$\varphi(\theta) = -\ell_\theta(\theta) \ . \tag{3.7}$$

A scalar rotated component of $\varphi(\theta)$ that agrees with (2.4) is then given by

$$\chi(\theta) = \varphi_p(\theta) = -\ell_\psi(\theta) \ .$$

Let $\lambda(\theta)$ be some nuisance parameterization that complements $\psi(\theta)$. The standardization in (2.6) is then given by

$$\frac{|\jmath_{(\theta\theta)}(\hat{\theta})|^{\frac{1}{2}}}{|\jmath_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{\frac{1}{2}}} = \frac{|\jmath_{\theta\theta}(\hat{\theta})|^{-\frac{1}{2}}}{|\jmath_{\lambda\lambda}(\hat{\theta}_\psi)|^{-\frac{1}{2}}}$$

using (3.7). We thus obtain (2.8) from (3.1).

The interesting aspect of the general formula applied to the Bayesian case is that we do not need to construct a nominal location model for (λ, ψ) but just for the initial parameterization θ.

4. EXAMPLES

For computer implementation we use Maple V to obtain the full and restricted maximum likelihood estimates, and to compute V, $\ell_{;V}$, $\ell_{\theta;V}$. It is often convenient to summarize the approximation of the left tail significance probability by a function $p(\psi)$ recording the probability to the left of the data in the frequentist case, where left indicates smaller maximum likelihood value, and probability to the right of ψ in the Bayesian case. In either case $p(\psi)$ decreases as ψ increases: it is a confidence distribution function in the frequentist case and is a survivor type function in the Bayesian case. This choice of left and right for the formula is conventional, and is trivially reversed by using $-\psi$ as the interest parameter. The Maple program used is available on `http://utstat.utoronto.ca/reid/research.html`.

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 $(326, 302, 307, 299, 329)$. Figure 1 shows the transcript of the Maple session used to generate the significance functions using (1.1) and (1.2): these are plotted in Figure 2. The ith pivotal quantity is simply $q_i = (y_i - \mu)/\sigma$. The 95% confidence interval using (1.1) or (1.2) is $(0.0267, 0.1281)$: Vangel's approximation gives the interval $(0.0270, 0.1293)$. Figure 2 also shows the significance function for the standard normal approximation to the signed square root of the log- likelihood ratio statistic, which in this case is not very accurate.

Figure 1: Maple V session for Example 4.1

```
with(linalg):
```

```
n:=5:
p:=2:
y0:=[326,302,307,299,329]:
sigma:=x[2]:
f[i]:=(1/((2*Pi)^ 0.5*sigma))*exp(-(y[i]-x[1])^ 2/(2*sigma^ 2));
f:=product(f[i],i=1..n);
g:=x[2]/x[1]:
q:=array([seq((y[i]-x[1])/x[2],i=1..n)]);
xhat:=[312.6,12.46755739]:
read cmle_file;
N:=70:
x0:=[ 313.0267219, 4.695400828, -2019.573491]:
for i from 1 to N do;
g0:=0.0125+0.0025*i;
b[i]:=xchat(20);
x0:=b[i]:
od;
read pvalue_file2;
for i from 1 to N do;
g0:=0.0125+0.0025*i;
xchat:=b[i];
alphahat:=xchat[p+1]:
pplr[i]:=plr();
ppLR[i]:=pLR();
ppBN[i]:=pBN();
od;
```

Figure 2 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 $\psi = \mu + (1/2)\sigma^2$, which is the logarithm of the mean of the associated log-normal distribution. Figure 3 shows the significance function for ψ for the data of Lieblein & Zelen (1956) using the normal approximation to r and using (1.1) and (1.2). Figure 4 shows the significance function for $\log \sigma$.

Figures 3 and 4 here

Example 4.3: Bayesian comparison of means

This example is taken from 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 $\bar{y}$ is the sample mean and S_i^2 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 to r and to r^*. 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). What is not as obvious from Hsu's histogram is that there is more weight in both tails of the third order approximation, compared to the standard normal approximation for r.

Table 4.1: nested posterior probability intervals

90%	1st order	$(0.105, 0.354)$
	3rd order	$(0.085, 0.368)$
95%	1st order	$(0.090, 0.395)$
	3rd order	$(0.067, 0.405)$
98%	1st order	$(0.075, 0.455)$
	3rd order	$(0.051, 0.475)$
99%	1st order	$(0.065, 0.505)$
	3rd order	$(0.042, 0.535)$

5. COMPARISON TO R^* WITH EXAMPLES

In this section we record an alternate form for the maximum likelihood departure Q of Fraser & Reid (1995), given here in (2.6). This enables explicit comparison with the expression u given in Barndorff-Nielsen (1986, 1991): this latter expression could also similarly be extended to the implicit parametrization case considered here in Section 3,

but for purposes of the comparison it is easier to restrict attention to the case of explicit nuisance and interest parameters.

By applying some elementary results on matrix determinants and definitions for the re-calibrated information determinants summarized in the Appendix, we obtain the following alternate expression for Q:

$$Q = \frac{|\ell_{;V}(\hat{\theta}) - \ell_{;V}(\hat{\theta}_\psi) \qquad \ell_{\lambda;V}(\hat{\theta}_\psi)|}{|\ell_{\theta;V}(\hat{\theta})|} \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{1/2}. \tag{4.1}$$

The expression for u from (3.3) of Barndorff-Nielsen (1991) or (6.108) of Barndorff-Nielsen and Cox (1994) has the form

$$u = \frac{|\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) \qquad \ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)|}{|\ell_{\theta;\hat{\theta}}(\hat{\theta})|} \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{1/2}. \tag{4.2}$$

For (4.2) it is assumed that the log-likelihood function can be expressed as $\ell(\theta; \hat{\theta}, a)$, where $\hat{\theta}$ is the maximum likelihood estimator, and differentiation with respect to $\hat{\theta}$ is for fixed a where a is an exactly or approximately ancillary statistic.

The equivalence is then obtained by noting that differentiation with respect to $\hat{\theta}$ given a is in fact differentiation with respect to vectors tangent to the ancillary and by using the equivalence $\ell_{\theta;\hat{\theta}}(\hat{\theta}) = j_{\theta\theta}(\hat{\theta})$; for this note that formula (4.1) is independent of the choice of tangent vectors.

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

$$Q = \frac{|\ell_{;t}(\hat{\theta}) - \ell_{;t}(\hat{\theta}_\psi) \qquad \ell_{\lambda;t}(\hat{\theta}_\psi)|}{|\ell_{\theta;t}(\hat{\theta})|} \left\{ \frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{1/2}. \tag{4.3}$$

As Q is invariant to one-to-one transformation of t, we have in this case that Q and u are identical.

In general we can conclude from (4.1) and (4.2) that for third order inference it is sufficient to compute the directional derivative of the log-likelihood function, and it is not

necessary to have an explicit expression for the exactly or approximately ancillary statistic a. This makes Q easier to calculate in many problems.

In the case of no nuisance parameters and thus a scalar parameter θ, (4.1) and (4.2) reduce respectively to

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

and

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

as noted in Reid (1996).

The connection between the general Q and u also gives alternate versions of Barndorff-Nielsen's (1993) modified profile likelihood:

$$L_{MP}(\psi) = |j_{\lambda\lambda}(\hat{\theta}_\psi)||\ell_{\lambda;V_\lambda}(\hat{\theta}_\psi)|^{-1}L(\hat{\theta}_\psi) \tag{4.4}$$

where $L(\hat{\theta}_\psi)$ is the profile likelihood function, and V_λ are the ancillary directions for the nuisance parameter λ. An alternative modification to the profile likelihood for ψ with third order properties was derived from (2.6) and recorded in Fraser & Reid (1996),

$$L_{Alt}(\psi) = |j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}L(\hat{\theta}_\psi)$$

where the nominal parameterization has been adjusted so that $j_{\phi\phi}(\hat{\theta}) = I$ is the identity. This is easily implemented using a symbolic computer program.

Example 5.1: 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

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

where $(t_1, t_2) = (\sum y_i^{-1}, \sum y_i)$ is the minimal sufficient statistic. The maximum likelihood estimates satisfy

$$(\hat{\psi}/\hat{\lambda})^{1/2} = t_2/n, \qquad \hat{\psi}^{-1} + (\hat{\lambda}/\hat{\psi})^{1/2} = t_1/n$$

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 we have from (4.3)

$$Q = (\hat{\psi} - \psi)(n\psi\hat{\psi}^{-3}/2)^{1/2}$$

which as expected is identical to that obtained by Barndorff-Nielsen (1990).

Example 5.2: Normal 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

$$\ell(\theta) = -n\log\theta - \frac{n}{2b^2\theta^2}t_2 + \frac{n}{b^2\theta}t_1$$

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

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

so that

$$Q = \{t_1(\hat{\theta}^{-1} - \theta^{-1}) - t_2(\hat{\theta}^{-2} - \theta^{-2})\}(\sqrt{n}\hat{\theta}/b)(t_2 + b^2\hat{\theta}^2)^{-1/2}.$$

There is an exact ancillary statistic for this model (Hinkley, 1977), which is $t_1/(t_2 - t_1^2)^{1/2}$, and the maximum likelihood estimate can be given explicitly as the root of a quadratic equation, so it would not be too 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. Table 5.1 below compares the exact and approximate values for selected values of b^2.

Table 5.1. Exact and approximate p-values: normal coefficient of variation.

$b^2 = 1$

μ	0.8	1	1.2	1.5	1.8	2	2.5	2.8
$\Phi(r)$	0.99755	0.93370	0.72726	0.37155	0.16431	0.09454	0.02576	0.01275
(1.2)	0.99855	0.95364	0.78600	0.44946	0.22169	0.13618	0.04262	0.02259
exact	0.99856	0.95403	0.78725	0.45124	0.22304	0.13717	0.04302	0.02282

$b^2 = 2$

μ	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
(1.2)	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

$b^2 = 1/2$

μ	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
(1.2)	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 (5.3): 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

$$f(y_1, y_2) = \chi\psi e^{-\psi} e^{-\chi y_1 - \psi y_2}. \tag{4.5}$$

Barndorff-Nielsen and Chamberlin (1991) considered third order inference for the submodel of (4.5) specified by $\chi = \chi(\psi) = \psi^{-1}e^{-\psi}$. Since no exact ancillary statistic exists for this model, they used the affine ancillary statistic to calculate u in (4.2). For the submodel, the pivotal quantity $q = (\chi(\psi)y_1, \psi(y_2 - 1))$ 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 have $\ell_{\psi;V}(\psi)=\chi(\psi)(1+\hat{\psi}^{-1})^2y_1+\hat{\psi}^{-1}(y_2-1)$. Combining this with the expression for the observed information we have

$$Q=\left\{(\frac{\chi(\psi)}{\chi(\hat{\psi})}-1)y_2+(\frac{\psi}{\hat{\psi}}-1)(y_2-1)\right\}(1+2\hat{\psi}^{-1}y_2-\hat{\psi}^{-1})^{-1}\{\chi(\hat{\psi})(1+2\hat{\psi}^{-1}+2\hat{\psi}^{-2})y_1\}^{\frac{1}{2}}. \tag{4.6}$$

Table 5.2 gives the results from 20,000 simulations of a sample of size 1 from the curved exponential model with $\psi=1$. We compare the empirical distribution function F_r of r, with the empirical distribution function F_u^* of r^* using u from Barndorff-Nielsen and Chamberlin (1991) and the empirical distribution function F_Q^* of r^* using Q in (4.6) to $N(0,1)$ distribution function at selected points.

Table 5.2. $N(0,1)$ distribution compared to the empirical cdf of r, r^* using u, and r^* using Q

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	0.176	0.0126	0.0007
F_u^*	0.0012	0.0229	0.236	0.485	0.258	0.0241	0.0013
F_Q^*	0.0012	0.0221	0.227	0.480	0.260	0.0241	0.0013
$\Phi(x)$	0.0013	0.025	0.251	0.500	0.251	0.025	0.0013

Example 5.4: Equality of gamma means

We assume we have 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

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

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 given as say $(y_{11},\ldots,y_{1n_1},y_{21},\ldots,y_{2n})$. A natural pivotal vector $q=(q_{11},\ldots,q_{1n_1},q_{21},\ldots,q_{2n})$ has entries $q_{ji}=F(y_{ji};\beta,\mu)$, where $F(y;\beta\mu)=F_\beta(\beta y/\mu)$ and $F_\beta(z)$ is the gamma (β) distribution function.

The matrix V needed to define φ is $(n_1+n_2)\times 3$:

$$\begin{pmatrix} \vdots & 0 & \vdots \\ \frac{\tilde{F}_{\beta_1}(y_{1i}/\mu)}{(\beta_1/\mu)f_{\beta_1}(\beta_1 y_{1i}/\mu)} & & \frac{y_{1i}}{\mu} \\ \vdots & \vdots & \vdots \\ 0 & \frac{\tilde{F}_{\beta_2}(y_{2i}/\mu)}{(\beta_2/\mu)f_{\beta_2}(\beta_2 y_{2i}/\mu)} & \frac{y_{2i}}{\mu} \\ \vdots & \vdots & \vdots \end{pmatrix}$$

where $\tilde{F}_\beta(\beta y/\mu) = \partial F(y;\beta,\mu)/\partial\beta$, and $f_\beta(y) = \Gamma^{-1}(\beta)y^{\beta-1}\exp(-y\beta)$. We can then compute $\ell_{;V}$, which is

$$\begin{pmatrix} a_1\beta_1 + b_1\beta_1/\mu \\ a_2\beta_2 + b_2\beta_2/\mu \\ (n_1/\hat{\mu})\beta_1 + (y_{1\cdot}/\hat{\mu})(\beta_1/\mu) + (n_2/\hat{\mu})\beta_2 + (y_{2\cdot}/\hat{\mu})(\beta_2/\mu) \end{pmatrix}$$

where $a_1 = a_1(\hat{\beta}_1,\hat{\mu}) = \sum^{n_1}\beta_1(y_{1i}^{-1}-\mu^{-1})c_1(\hat{\beta}_1,\hat{\mu},y_{1i})$, and $c_1(\beta_1,\mu,y_{1i})$ is the $(i,1)$ entry of V, and $b_1 = \sum_i c_1(\beta_1,\mu,y_{1i})$. These are readily combined into expression (4.1) above.

APPENDIX: Derivation of (4.2)

When ψ is an explicit component of θ, which we here write as (ψ,λ) the vector ψ_θ needed in (2.5) is $(1,0,\cdots,0)$. Using the following three matrix identities,

$$(1,0,\cdots,0)\begin{pmatrix} a & \alpha^T \\ \beta & B \end{pmatrix}\begin{pmatrix} b \\ \gamma \end{pmatrix} = \begin{vmatrix} b & \alpha^T \\ \gamma & B \end{vmatrix} \Big/ \begin{vmatrix} a & \alpha^T \\ \beta & B \end{vmatrix}$$

$$|(1,0,\cdots,0)\begin{pmatrix} a & \alpha^T \\ \beta & B \end{pmatrix}^{-1}| = |B|(1+\alpha^T B^{-2}\alpha)^{1/2} \Big/ \left|\begin{pmatrix} a & \alpha^T \\ \beta & B \end{pmatrix}\right|$$

$$|(\,\alpha \qquad B^T\,)\begin{pmatrix} \alpha^T \\ B \end{pmatrix}| = |B|^2(1+\alpha^T B^{-2}\alpha),$$

we obtain the following results:

$$\begin{aligned} \psi_\varphi(\hat{\theta}_\psi)\{\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)\} &= \psi_\theta \cdot \varphi_\theta^{-1}(\hat{\theta}_\psi)\{\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)\} \\ &= |\ell_{;V}(\hat{\theta}) - \ell_{;V}(\hat{\theta}_\psi) \quad \ell_{(\lambda;V}(\hat{\theta}_\psi)|/|\varphi_\theta(\hat{\theta}_\psi)|, \\ ||\psi_\varphi(\hat{\theta}_\psi)|| &= |\varphi_\lambda^T(\hat{\theta}_\psi)\varphi_\lambda(\hat{\theta}_\psi)|^{1/2}/|\varphi_\theta(\hat{\theta}_\psi)|, \end{aligned}$$

and

$$|j_{(\lambda\lambda)}(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi_\lambda^T(\hat{\theta}_\psi)\varphi_\lambda(\hat{\theta}_\psi)|^{-1},$$

$$|j_{(\theta\theta)}(\hat{\theta})| = |j_{\theta\theta}(\hat{\theta})||\varphi_\theta(\hat{\theta})|^{-2} = |j_{\theta\theta}(\hat{\theta})||\ell_{\theta;V}(\hat{\theta})|^{-2}$$

Inserting these into (2.6) gives (4.2)

REFERENCES

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**, 485–96.

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

Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and saddlepoint approximations with statistical inference. *J.R. Statist. Soc. B* **41**, 279-312.

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

Cheah, P.K., Fraser, D.A.S., and Reid, N. (1995). Adjustment to likelihood and densities; calculating significance. *Journal of Statistical Research* **29**, 1-13.

DiCiccio, T., and 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.

Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65-76.

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

Fraser, D.A.S. and Reid, N. (1996). Bayes posteriors for scalar interest parameters. *Bayesian Statistics,* **V**, 581-585.

Hinkley, D.V. (1977). Conditional inference about a normal mean with known coefficient of variation. *Biometrika* **64**, 105–8.

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

Lieblein, J. and 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. and Rice, S.O. (1980). Saddlepoint approximation for the distribution of the sums of independent random variables. *Adv. Appl. Prob.* **12**, 475-490.

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

Figure 2: Significance probability for normal coefficient of variation (Example 4.1). Solid line: normal approximation to r; Dotted line: approximation (1.1); Dashed line: approximation (1.2).

Figure 3: Significance probability for log-normal mean (Example 4.2). Solid line: normal approximation to r; Dotted and dashed lines: approximations (1.1) and (1.2).

Figure 4: Significance probability for log-normal variance (Example 4.3). Solid line: normal approximation to r; Dotted and dashed lines: approximations (1.1) and (1.2).

Figure 2

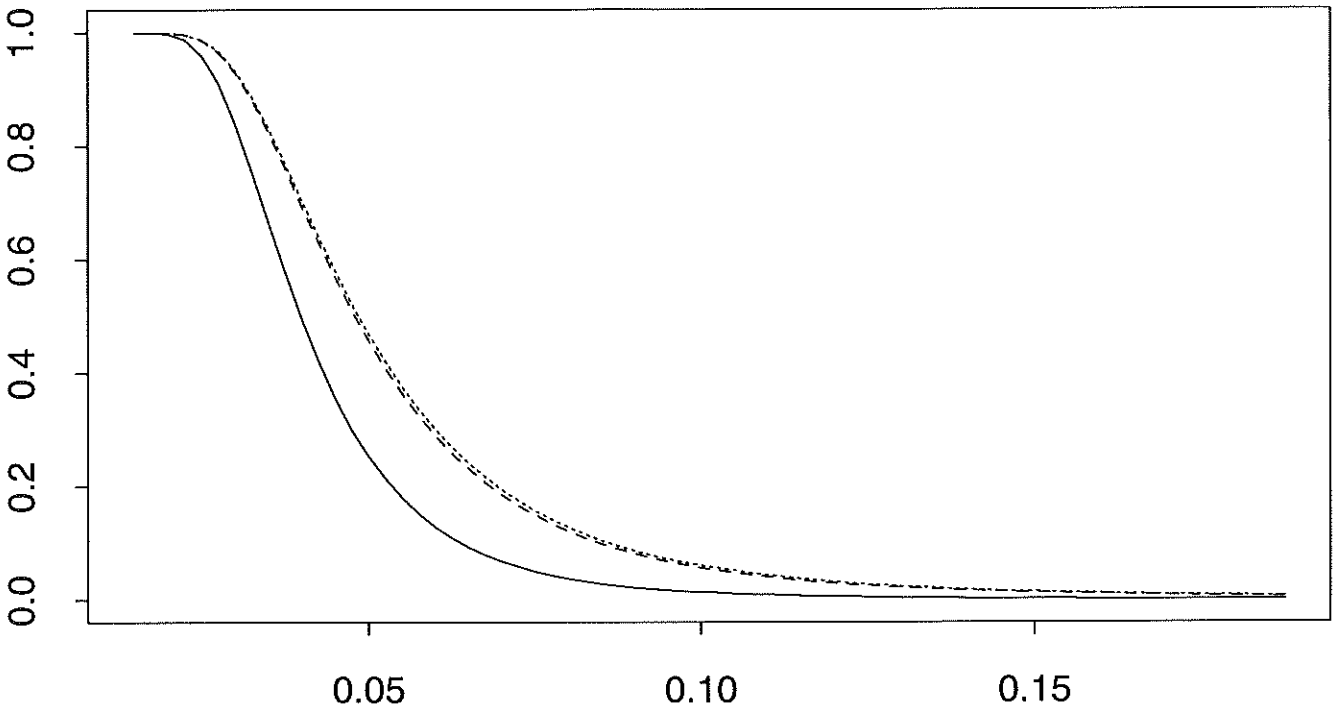

Figure 3

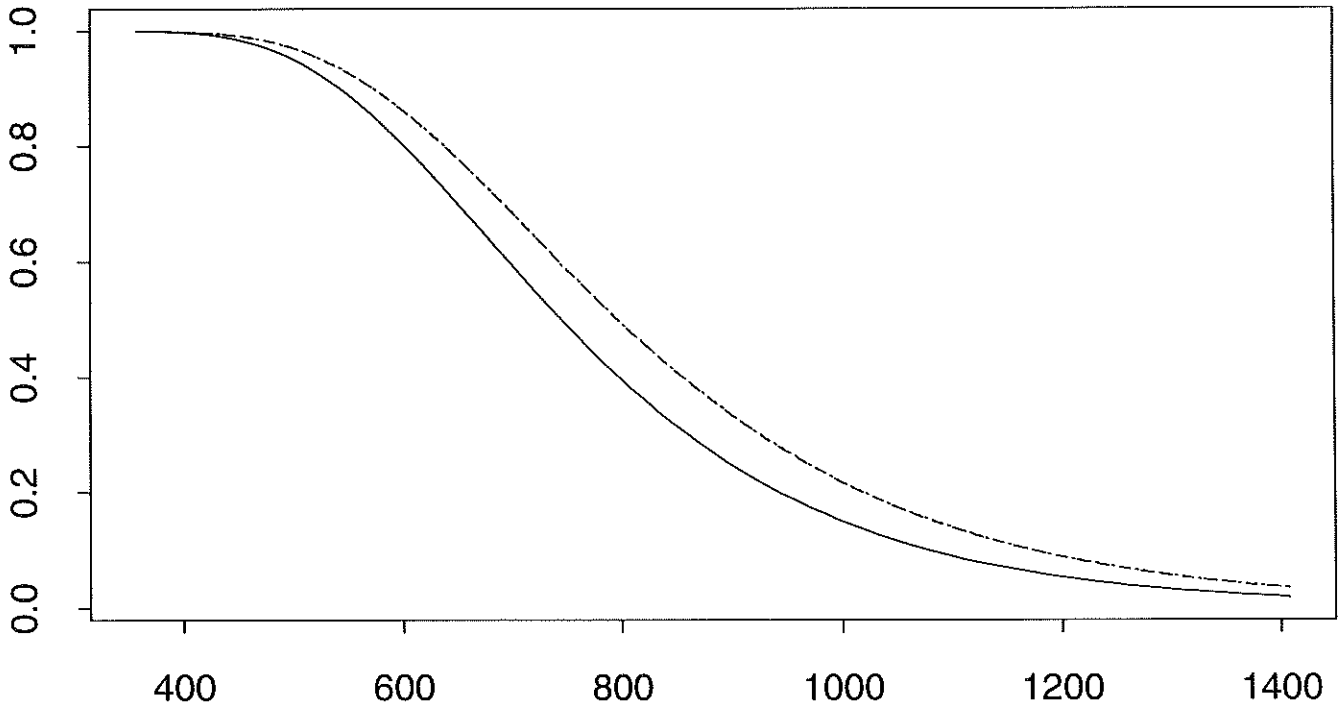

Figure 4