Theoretical Statistics and Asymptotics

Nancy Reid

University of Toronto
Department of Statistics
100 St. George St., Toronto
Canada

Email:reid@utstat.utoronto.ca

1 Introduction

Cox and Hinkley's *Theoretical Statistics* (1974) was arguably the first modern treatment of the foundations and theory of statistics, and for the past thirty years has served as a unique book for the study of what used to be called mathematical statistics. One of its strengths is the large number of examples that serve not only to illustrate key concepts, but also to illustrate limitations of many of the classical approaches to inference. Since the time of its publication there has been considerable development of the theory of inference, especially in the area of higher order likelihood asymptotics. The paper by Barndorff-Nielsen and Cox (1979) read to the Royal Statistical Society initiated many of these developments. There have also been very dramatic advances in the implementation of Bayesian methodology, particularly with the development of sampling methods for computing posterior distributions.

A current approach to the study of theoretical statistics should be informed by these advances. Barndorff-Nielsen and Cox (1994) presents an approach to inference based on likelihood asymptotics, as does Pace and Salvan (1997). Severini (2000a) covers much of the same material at a somewhat more elementary level. Bayesian methodology is covered in books on that topic, for example Bernardo and Smith (2000), while Schervish (1995) presents an encyclopedic treatment of both classical and Bayesian approaches to inference. Davison (2003) provides a masterful and wide-ranging account of the use of statistical models in theory and practice, with a welcome emphasis on many practical aspects, and this approach is closest in spirit to the one I would like to follow. At a more advanced level there is not to my knowledge a treatment of the main themes of theoretical statistics that attempts to incorporate recent advances in a unified framework.

In this paper I will attempt to set out some ideas for such a framework, building on topics and examples from Cox and Hinkley (1974), and very heavily influenced by Barndorff-Nielsen and Cox (1994). The emphasis is on some theoretical principles that have their basis in asymptotics based on the likelihood function. There are of course many important advances and developments in statistical theory and methodology that are not subsumed by asymptotic theory or Bayesian inference. These include developments in nonparametric and semiparametric inference as well as a large number of more specialised techniques for particular applications. The explosion in statistical methods in a great many areas of science have made a comprehensive study of the field of statistics much more difficult than perhaps it was, or seems to have been, thirty years ago. In light of David Cox's view that a theory of statistics forms a basis both for understanding new problems and for developing new methods, I try here to focus on how to structure this theory for students of statistics.

In §2 I sketch a framework for the implementation of higher order asymptotics and Bayesian inference into my ideal course or text on the theory of statistics. In §3 I note topics from a more standard treatment of theoretical statistics that have been omitted and provide some brief conclusions.

2 Likelihood based asymptotics

2.1 Introduction

The likelihood function provides the foundation for the study of theoretical statistics, and for the development of statistical methodology in a wide range of applications. The basic ideas for likelihood are outlined in many books, including Ch. 2 of Cox and Hinkley (1974). Given a parametric model $f(y;\theta)$ on a suitably defined sample space and parameter space, the likelihood and log-likelihood functions of θ are defined as

$$L(\theta) = L(\theta; y) \propto f(y;\theta) \qquad \ell(\theta) = \ell(\theta; y) = \log L(\theta; y) + a(y).$$

If θ is a scalar parameter a plot of $L(\theta)$ or $\ell(\theta)$ against θ for fixed y provides useful qualitative information about the values of θ consistent with that observed value of y. For examples see Barndorff-Nielsen and Cox (1994, pp. 96–97). Particularly useful summaries include the score function $u(\theta; y) = \ell'(\theta)$, the maximum likelihood estimate $\hat\theta = \arg\sup_\theta \ell(\theta)$, the observed Fisher information $j(\theta) = -\ell''(\theta)$ and the log-likelihood ratio $\ell(\hat\theta) - \ell(\theta)$. Standardized versions of these are used to measure the discrepancy between the maximum likelihood estimate $\hat\theta$ and the parameter θ: we define them as follows:

$$\begin{aligned} \text{score} \qquad s(\theta) &= \ell'(\theta)\{j(\hat\theta)\}^{-1/2} && (1)\\ \text{Wald} \qquad q(\theta) &= (\hat\theta - \theta)\{j(\hat\theta)\}^{1/2} && (2)\\ \text{likelihood root} \qquad r(\theta) &= \text{sign}\ (q)\sqrt{[2\{\ell(\hat\theta) - \ell(\theta)\}]}. && (3) \end{aligned}$$

With independent identically distributed observations $y = (y_1, \ldots, y_n)$, an application of the central limit theorem to the score function $u(\theta; y)$ can be used to show each of s, q and r has a limiting standard normal distribution, under the model $f(y;\theta)$, and subject to conditions on f that ensure consistency of $\hat\theta$ and sufficient smoothness to permit the requisite Taylor series expansions.

These limiting results suggest that in finite samples we can approximate the distribution of s, q, or r by the standard normal distribution. In Figure 1 we illustrate these approximations for a single observation from the location model $f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}$. I have plotted as a function of θ the exact probability $F(y;\theta)$ and the normal approximations to this exact probability: $\Phi(s)$, $\Phi(q)$ and $\Phi(r)$, for a fixed value of y, taken here to be 21.5. This is called here the p-value function, and is used to calibrate the log-likelihood function by identifying values of θ consistent with the data on a probability scale. For example, the value of θ that corresponds to a p-value of $1-\alpha$ corresponds to a lower $100\alpha\%$ confidence bound for θ. The approximations illustrated in Figure 1 are called *first order* approximations, as they are based on the limiting distribution.

A function of y and θ that has a known distribution is often called a *pivotal* statistic, or pivotal function. The p-value functions based on r, q and s shown in Figure are all pivotal functions, at least approximately, since the distribution of, for example, r, is known to be standard normal (approximately), and so the distribution of $\Phi(r)$ is $U(0,1)$ (approximately). The exact p-value function constructed from the known distribution $f(y;\theta)$ is an exact pivotal function. An important aspect of the theory of higher order asymptotics based on the likelihood function is the joint dependence of the likelihood function and derived quantities on both the parameter and the data.

Somewhat remarkably, a simple combination of r and s leads to a pivotal function for which the normal approximation is much improved more accurate. Table 1 compares the exact p-value to that based on a normal approximation to

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log \frac{s(\theta)}{r(\theta)}. \qquad (4)$$

The derivation of r^* uses so-called higher order asymptotics, the basis of which is an asymptotic expansion whose leading term gives the first order normal approximation to the distribution of r. The approximation illustrated in Table 1 is often called a *third order* approximation, as detailed analysis shows that the approximation has relative error $O(n^{-3/2})$, in the *moderate deviation* range $\hat\theta - \theta = O_p(n^{-1/2})$. However, before theoretical motivation is provided for this particular construction the numerical evidence is compelling, in this and any number of continuous one-parameter models. The next subsections consider methods for obtaining similar results in somewhat more complex models. The emphasis is on constructing p-value functions for a scalar parameter, based on a scalar summary of the data. Subsection 2.2 considers reduction of the data to a scalar summary, and subsection 3.3 discusses models with nuisance parameters.

Figure 1: The exact distribution of $\hat{\theta}$ (solid) and the approximate normal distribution based on the likelihood root r (dash), the score statistic s (dotted) and the Wald statistic q (long dash). The model is $f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}$ and the observed value of y is 21.5. This illustrates the use of asymptotic theory to give approximate p-values; in this case the approximations are quite different from each other and from the exact p-value.

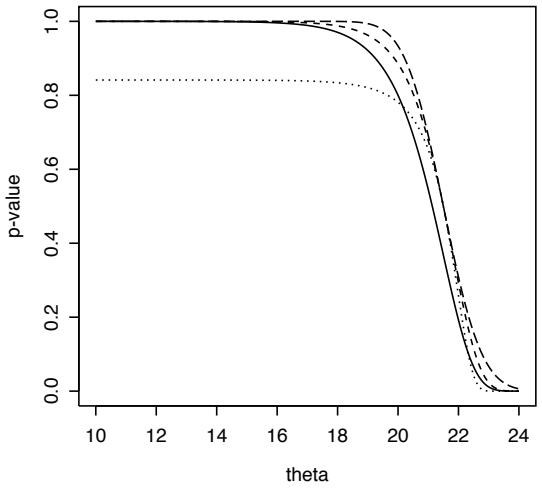


Table 1: Selected exact p-values compared to the third order approximation $\Phi(r^*)$, for the location model illustrated in Figure 1.

θ	14.59	16.20	16.90	17.82	18.53	20.00	20.47	20.83
exact	0.999	0.995	0.990	0.975	0.950	0.800	0.700	0.600
$\Phi(r^*)$	0.999	0.995	0.990	0.974	0.949	0.799	0.700	0.600
θ	21.41	21.69	21.98	22.60	22.81	23.03	23.17	23.43
exact	0.400	0.300	0.200	0.050	0.025	0.010	0.005	0.001
$\Phi(r^*)$	0.401	0.302	0.202	0.051	0.025	0.010	0.005	0.001

2.2 Marginal and conditional distributions

With a sample of size 1 and a scalar parameter the variable is in one-to-one correspondence with the parameter, and plots and tables are easily obtained. In some classes of models we can establish such a correspondence and obtain analogous results. In full exponential families this correspondence is achieved by considering the marginal distribution of the maximum likelihood estimator or the canonical statistic. In transformation families this correspondence is achieved by considering the conditional distribution given the maximal ancillary. Thus these two families can serve as prototypes for proceeding more generally. The discussion of exponential and transformation families in Chapter 2 of Barndorff-Nielsen and Cox (1994) seems ideal for this, and there are excellent treatments as well in Pace and Salvan (1997), Severini (2000a) and Davison (2003). An important and useful result is the derivation, in transformation models, of the exact distribution of the maximum likelihood estimator conditional on the maximal ancillary. The illustration of this is particularly simple for the location model, where we have (Cox and Hinkley, 1974, Ex. 4.15),

$$f(\hat{\theta} \mid a; \theta) = \frac{\exp\{\ell(\theta; \hat{\theta}, a)\}}{\int \exp\{\ell(\theta; \hat{\theta}, a)\} d\theta} \tag{5}$$

with

$$\ell(\theta; \hat{\theta}, a) = \sum \log f(y_i - \theta) = \sum \log f(a_i + \hat{\theta} - \theta), \tag{6}$$

where the maximal ancillary $a = (a_1, \ldots, a_n) = (y_1 - \hat{\theta}, \ldots, y_n - \hat{\theta})$.

Laplace approximation to the integral in the denominator gives an approximation to the conditional density of $\hat{\theta}$, called the p^* approximation

$$p^*(\hat{\theta} \mid a; \theta) = c|j(\hat{\theta})|^{1/2} \exp\{\ell(\hat{\theta}) - \ell(\theta)\}. \tag{7}$$

An approximation to the integral of this density can be integrated to give the normal approximation to the distribution of r^* defined at (4); see for example Davison (2003, Ch.12). This argument generalises to transformation families, of which the location model is the simplest example (Barndorff-Nielsen and Cox, 1994, §6.2).

Related calculations can be carried out in one parameter exponential families with density $f(y_i; \theta) = \exp\{\theta y_i - c(\theta) - d(y_i)\}$. For $y = (y_1, \ldots, y_n)$ we have

$$f(y; \theta) = \exp\{\theta y_+ - nc(\theta) - \tilde{d}(y_+)\};$$

where $y_+ = \Sigma y_i$ is a minimal sufficient statistic for θ. A higher order approximation to the distribution of y_+ or $\hat{\theta}$ leads to a normal approximation to the distribution of

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log \frac{q(\theta)}{r(\theta)} \tag{8}$$

where now $q(\theta)$ is the Wald statistic (2). This approximation can be derived from a saddlepoint approximation to the density of the sufficient statistic y_+, just as (4) is derived by integrating the p^* approximation (7). An example is given by the Pareto distribution

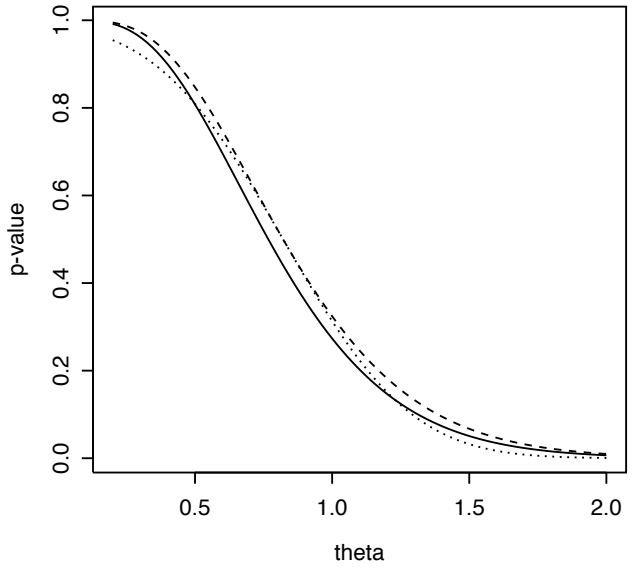
$$f(y; \theta) = \theta(1 + y)^{-(\theta+1)}, \quad y > 0, \quad \theta > 0;$$

the p-values based on a simulated sample of size 5 are illustrated in Figure 2. The exact distribution is computed using the fact that $\log(1 + y)$ follows an exponential distribution with rate θ.

Examples such as these are perhaps useful for motivation of the study of the r^* approximation, but they are also essentially defining the inference goal as a set of confidence bounds, or equivalently p-values for testing arbitrary values of the parameter. The intervals will not be nested if the log-likelihood function is multi-modal, so it is important to plot the likelihood function as well as the p-value function.

In location families, or more generally in transformation families, conditioning serves to reduce the dimension of the data to the dimension of the parameter. Assume we have a model with a single unknown parameter, that is neither an exponential family model nor a transformation model. The solution that arises directly from consideration of higher order approximations is to construct a one-dimensional conditional model using an *approximate* ancillary statistic.

Figure 2: A one-parameter exponential family illustration of the asymptotic theory. The approximate p-value based on r^* cannot be distinguished from the exact value (solid). The normal approximation to the likelihood root (dashed) and Wald statistic (dotted) are much less accurate.



As an example consider sampling from the bivariate normal distribution, with only the correlation coefficient as an unknown parameter (Cox and Hinkley, 1974, Ex. 2.30; Barndorff-Nielsen and Cox, 1994, Ex. 2.38). The log-likelihood function is

$$\ell(\theta) = -\frac{n}{2}\log(1-\theta^2) - \frac{1}{2(1-\theta^2)}\sum(y_{1i}^2 + y_{2i}^2) + \frac{\theta}{1-\theta^2}\sum y_{1i}y_{2i}, \qquad -1 \le \theta \le 1. \tag{9}$$

In this model the minimal sufficient statistic $(\Sigma(y_{1i}^2 + y_{2i}^2), \Sigma y_{1i}y_{2i})$ does not have a single component that is exactly ancillary, although either of Σy_{1i}^2 or Σy_{2i}^2 is ancillary. It is possible to find an approximate ancillary by embedding the model in a two-parameter exponential family, for example by treating $\alpha_1\Sigma(y_{1i}+y_{2i})^2/\{2n(1+\theta)\}$ and $\alpha_2\Sigma(y_{1i}-y_{2i})^2/\{2n(1-\theta)\}$ as independent χ_n^2 random variables, where $\alpha_1 = \alpha/(1-\theta)$, $\alpha_2 = \alpha/(1+\theta)$, thus recovering (9) when $\alpha = 1$. An illustration of this approach in a different (2,1) family is given in Barndorff-Nielsen and Cox (1994, Ex. 7.1).

It is also possible to use a method based on a local location family, as described in Fraser and Reid (1995). The main idea is to define the approximately ancillary statistic indirectly, rather than explicitly, by finding vectors in the sample space tangent to an approximate ancillary. This is sufficient to provide an approximation to the p-value function. The details for this example are given in Reid (2003, §4), and the usual normal approximation to the distribution of r^* still applies, where now

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)}\log\frac{Q(\theta)}{r(\theta)} \tag{10}$$

and

$$Q(\theta) = \{\varphi(\hat\theta) - \varphi(\theta)\}\{j_{\varphi\varphi}(\hat\theta)\}^{1/2} \tag{11}$$

is the Wald statistic (2) in a new parametrisation $\varphi = \varphi(\theta; y)$. This new parametrisation is the canonical parameter in an approximating exponential model and is computed by sample space differentiation of the log-likelihood function:

Table 2: Exact and approximate p-values for various values of θ. The exact values are based on 100,000 simulations from the model given at (9), with $n = 5$. The second row is based on the normal approximation to the distribution of r, and the third row is based on the normal approximation to the distribution of r^*.

		left tail					right tail				
θ	Exact	0.001	0.005	0.010	0.025	0.050	0.050	0.025	0.010	0.005	0.001
-0.9	r	0.0022	0.009	0.018	0.042	0.079	0.033	0.016	0.006	0.003	0.0005
	r^*	0.0010	0.005	0.010	0.025	0.050	0.050	0.025	0.010	0.005	0.0009
-0.7	r	0.0019	0.009	0.018	0.041	0.078	0.045	0.024	0.011	0.006	0.0015
	r^*	0.0009	0.005	0.010	0.025	0.049	0.053	0.027	0.011	0.005	0.0013
-0.5	r	0.0022	0.009	0.017	0.041	0.077	0.058	0.031	0.014	0.007	0.0017
	r^*	0.0010	0.005	0.010	0.025	0.050	0.055	0.027	0.010	0.005	0.0010
-0.3	r	0.0022	0.009	0.019	0.042	0.076	0.065	0.035	0.016	0.009	0.0019
	r^*	0.0011	0.005	0.010	0.026	0.051	0.053	0.026	0.011	0.005	0.0010
0	r	0.0022	0.010	0.018	0.040	0.074	0.072	0.039	0.017	0.009	0.0021
	r^*	0.0011	0.005	0.011	0.026	0.051	0.051	0.025	0.010	0.005	0.0010

$$\begin{aligned} \varphi(\theta) &= \ell_{;V}(\theta; y), \\ j_{\varphi\varphi}(\hat\theta) &= |\ell_{\theta;V}(\hat\theta)|^{-2} j(\hat\theta). \end{aligned} \tag{12}$$

The derivation of V is outlined in Reid (2003, §3); briefly if $z_i(y_i;\theta)$ has a known distribution, then the ith element of V is

$$V_i = -(\frac{\partial z_i}{\partial y_i})^{-1}(\frac{\partial z_i}{\partial \theta})|_{\hat\theta}. \tag{13}$$

For model (9) we have

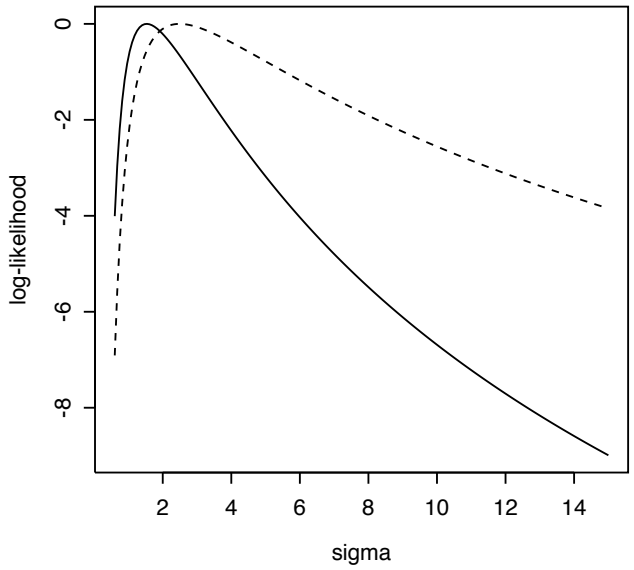
$$\begin{aligned} \ell_{;V}(\theta) &= n\{\theta(t - \hat\theta s) - (s - \hat\theta t)\}/\{(1-\theta^2)(1-\hat\theta^2)\} \\ \ell_{\theta;V}(\hat\theta) &= n(t - 3\hat\theta s + 3\hat\theta^2 t - \hat\theta^3 s)/\{(1-\hat\theta^2)^3\} \end{aligned}$$

with $s = \Sigma(y_{1i}y_{2i})$, $t = \Sigma(y_{1i}^2 + y_{2i}^2)$. The pivotal statistic used was $(z_{1i}, z_{2i}) = [(y_{1i} + y_{2i})^2/\{2(1+\theta)\}, (y_{1i} - y_{2i})^2/\{2(1-\theta)\}$.

Table 2shows exact and approximate p-values for inference about θ based on the normal approximation to r and to r^*. The exact values were obtained from 100,000 simulations. The approximation is remarkably accurate even for extreme values of θ.

The derivations of approximations like these require a fairly detailed description of asymptotic theory, both the usual first order theory and less familiar extensions such as Edgeworth, saddlepoint, and Laplace approximations, as set out, for example, in Barndorff-Nielsen and Cox (1990). It is difficult to know how much of this is essential for a modern approach to statistical theory, but one hopes it is possible to describe the main results without being overly technical or overly vague. I have not been specific about the definition of *approximate ancillarity*, which can get rather technical. The construction using vectors based on a pivotal statistic gives a means of proceeding, and while the ancillary statistic related to this is not uniquely determined, the resulting approximation is unique to third order.

Figure 3: The profile log-likelihood (solid) and the marginal log-likelihood (dashed) for σ in the one-way analysis of variance model, with three observations from each of five groups. Profiling does not properly account for uncertainty in the estimation of the nuisance parameter λ.



2.3 Models with nuisance parameters

We now consider vector parameters θ, and in order to be able to extend the construction of a p-value function we will assume that the parameter of interest ψ is a scalar, and that $\theta = (\psi, \lambda)$. The simplest way to obtain a log-likelihood function for ψ is to use the profile log-likelihood function $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ where $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimate. Properties and illustrations of the profile log-likelihood function are described in Barndorff-Nielsen and Cox (1994, Ch. 3.4,5). Asymptotically normal statistics analogous to (1), (2) and (3) can be obtained from the profile log-likelihood.

The profile likelihood does not in general correspond to the density of an observable random variable, however, and as a result the first order approximations can be very poor, especially in the case of high dimensional nuisance parameters, such as inference for the variance parameter in a normal theory linear regression, and in more extreme examples where the dimension of the nuisance parameter increases with the sample size. These examples can be used to motivate adjustments to the profile likelihood to accommodate the presence of nuisance parameters.

As an example consider a one-way analysis of variance $y_{ij} \sim N(\lambda_i, \psi), i = 1, \ldots n; j = 1, \ldots, k$ (Cox and Hinkley, 1974, Ex. 5.20, 9.5, 9.24). A comparison of the profile log-likelihood and the log-likelihood from the marginal distribution of the within group sum of squares $\Sigma_{ij}(y_{ij} - \bar{y}_{i.})^2$ illustrates both the effect on the maximum point and on the curvature; see Figure 3. The marginal log-likelihood for variance components in normal theory linear models is often described as a REML log-likelihood, and has been discussed in much more general settings; see, for example, Cox and Solomon (2002, Ch. 4).

I think the easiest way to motivate an the adjustment that in this example produces the marginal log-likelihood

is to take a Bayesian approach, and derive the Laplace approximation to the marginal posterior density

$$\begin{aligned}\pi_m(\psi \mid y) &= \frac{\int \exp\{\ell(\psi,\lambda;y)\pi(\psi,\lambda)d\lambda}{\int \exp\{\ell(\psi,\lambda;y)\pi(\psi,\lambda)d\psi d\lambda} \\ &\doteq \frac{1}{\sqrt{(2\pi)}}\exp\{\ell_{\mathrm{p}}(\psi)-\ell_{\mathrm{p}}(\hat\psi)\}\{j_{\mathrm{p}}(\hat\psi)\}^{1/2}\frac{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)|^{1/2}}{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}\frac{\pi(\psi,\hat\lambda_\psi)}{\pi(\hat\psi,\hat\lambda)} \qquad (14)\end{aligned}$$

where $j_{\mathrm{p}}(\psi) = -\ell''_{\mathrm{p}}(\psi)$, and $j_{\lambda\lambda}(\psi,\lambda)$ is the nuisance parameter component of the observed Fisher information matrix.

This approximation is relatively easy to derive, and can be used to develop two important and closely related aspects of higher order asymptotics in a frequentist setting. First, the 'likelihood function' associated with this marginal posterior is an adjusted profile log-likelihood

$$\ell_{\mathrm{a}}(\psi) = \ell_{\mathrm{p}}(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|.$$

It seems at least plausible that one could construct non-Bayesian inference for ψ by ignoring the prior, i.e. assigning a uniform prior to both ψ and λ. This shows directly that the adjusted log-likelihood is not invariant to reparameterisation, and could lead into a discussion of orthogonal parameterisation (Barndorff-Nielsen and Cox, 1994, Ch. 2.7; Cox and Reid, 1987).

Second, an extension of the derivation of the r^* approximation in the location model leads to an approximation to the marginal posterior survivor function of the same form

$$\int_\psi^\infty \pi_m(\psi \mid y)d\psi \doteq \Phi(r^*) \qquad (15)$$

where

$$\begin{aligned}r^* = r^*(\psi) &= r_{\mathrm{p}}(\psi) + \frac{1}{r_{\mathrm{p}}(\psi)}\log\frac{q_B(\psi)}{r_{\mathrm{p}}(\psi)} \\ r_{\mathrm{p}}(\psi) &= \text{sign }\{q_B(\psi)\}\sqrt{[2\{\ell_{\mathrm{p}}(\hat\psi)-\ell_{\mathrm{p}}(\psi)\}]} \\ q_B(\psi) &= -\ell'_{\mathrm{p}}(\psi)\{j_{\mathrm{p}}(\hat\psi)\}^{-1/2}\frac{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)|^{1/2}}\frac{\pi(\hat\psi,\hat\lambda)}{\pi(\psi,\hat\lambda_\psi)}. \qquad (16)\end{aligned}$$

In a nonBayesian context, higher order inference in the presence of nuisance parameters is relatively straightforward in two models that, while specialised, never-the-less cover a fair number of examples. The first is independent sampling from the regression-scale model

$$y_i = x_i'\beta + \sigma e_i \qquad (17)$$

where e_i are independent, mean 0, and follow a known distribution. A generalisation of the location family result (5) gives a density on R^{p+1}, where p is the dimension of β, by conditioning on the residuals $(y_i - x_i'\hat\beta)/\hat\sigma$. In this conditional distribution all nuisance parameters are exactly eliminated by marginalizing to the t-statistic for a component of β (or for $\log\sigma$);

$$t_j = (\hat\beta_j - \hat\beta)/\hat{se}(\hat\beta_j).$$

This is essentially the same as Bayesian inference for β_j using the prior $d\beta d\log\sigma$, so the approximation given at (15) can be used. Software for this is implemented in the `marg` library of `S` (Brazzale, 2000, Ch. 6). Table 3gives 95% confidence intervals for a regression coefficient fitting model (17) with a number of different error distributions. It is computed from the house price data from Sen and Srivastava (1990). In this example the confidence intervals are quite stable over a range of error distributions, although the first- and third- order confidence intervals are rather different.

In exponential family models, if the parameter of interest is a component of the canonical parameter, then the nuisance parameter can be eliminated by conditioning on the remaining components of the sufficient statistic, leading to approximations like (15) for models such as logistic regression, Poisson regression, and inference about the shape

Table 3: Comparison of 95% confidence intervals for a regression parameter under different models for the error distribution. Fitting and inference is carried out in R using the `marg` library of Brazzale. The data set has 26 observations, 4 covariates and an unknown scale parameter. The confidence intervals are for the coefficient of the covariate `frontage`.

	First order		Third order	
Student (3)	-0.07	0.65	-0.16	0.69
Student (5)	-0.09	0.65	-0.15	0.70
Student (7)	-0.08	0.66	-0.14	0.70
Normal	-0.08	0.66	-0.13	0.71

of a Gamma distribution. The appropriate definition of r^* can be derived from the saddlepoint approximation as

$$\begin{aligned} r^* = r^*(\psi) &= r_{\mathrm{p}}(\psi) + \frac{1}{r_{\mathrm{p}}(\psi)} \log \frac{q_E(\psi)}{r_{\mathrm{p}}(\psi)} \\ r_{\mathrm{p}}(\psi) &= \text{sign } (\hat\psi - \psi)\sqrt{[2\{\ell_{\mathrm{p}}(\hat\psi) - \ell_{\mathrm{p}}(\psi)\}]} \\ q_E(\psi) &= (\hat\psi - \psi)\{j_{\mathrm{p}}(\hat\psi)\}^{1/2} \frac{|j_{\lambda\lambda}(\hat\psi, \hat\lambda)|^{1/2}}{|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|^{1/2}}. \end{aligned} \tag{18}$$

This is implemented in `cond` in Brazzale (2000), where a number of examples are given. In regression models for discrete data, the question of continuity correction arises, and the correct interpretation of the third order approximations, which are inherently continuous, is not completely clear. Experience seems to indicate that the normal approximation to the distribution of r^* gives good approximations to the mid-p-value, i.e. $\Pr(y > y^0) + (1/2)\Pr(y = y^0)$. The continuity correction implemented in Brazzale (2000) instead replaces y^0 by $y^0 \pm 1/2$, according as y^0 is in the right or left tail, and computes r^* at this new data value. Discussion of continuity correction is given in Davison and Wang (2002), Pierce and Peters (1992) and Severini (2000b).

Outside these two classes there is not an exact conditional or marginal density for the parameter of interest. To find an approximate solution, it turns out that the higher order approximation is a generalisation of the t-pivot result for regression-scale models. The approach developed by Barndorff-Nielsen (1986) and outlined in Barndorff-Nielsen and Cox (1994, Ch.6.6) is based on an approximation to the conditional density of $\hat\theta$ given a, which is transformed to the joint density of $(r^*(\psi), \hat\lambda_\psi)$, where now

$$r^* = r^*(\psi) = r_{\mathrm{p}}(\psi) + \frac{1}{r_{\mathrm{p}}(\psi)} \log \frac{u(\psi)}{r_{\mathrm{p}}(\psi)}$$

and to show that this leads to the standard normal approximation to the density of r^*. The complementing statistic $u(\psi)$ is

$$u(\psi) = |\ell_{;\hat\theta}(\hat\theta) - \ell_{;\hat\theta}(\hat\theta_\psi) \quad \ell_{\lambda;\hat\theta}(\hat\theta_\psi)| / \{|j_{\lambda\lambda}(\hat\theta_\psi)||j(\hat\theta)|\}^{1/2}; \tag{19}$$

this assumes that the log-likelihood $\ell(\theta)$ can be expressed as $\ell(\theta; \hat\theta, a)$. It can in applications be quite difficult to compute the sample space derivatives in u, and a method of approximating them when the model can be embedded in a full exponential family is proposed in Skovgaard (1996, 2001).

An approach developed in Fraser (1990) and Fraser and Reid (1995), and outlined in Reid (2003) uses the parametrisation $\varphi(\theta)$ given at (12) to generalise Q of (11) to

$$\begin{aligned} Q(\psi) &= \{\nu(\hat\theta) - \nu(\hat\theta_\psi)\}/\hat\sigma_\nu \\ &= \frac{|\varphi(\hat\theta) - \varphi(\hat\theta_\psi) \quad \varphi_{\lambda'}(\hat\theta_\psi)|}{|\varphi_{\theta'}(\hat\theta)|} \frac{|j_{\theta\theta}(\hat\theta)|^{1/2}}{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}}. \end{aligned} \tag{20}$$

The close connection between u and Q is discussed further in Reid (2003) and in Fraser, Reid and Wu (1999). Both ν and $\hat\sigma_\nu$ are computed directly from $\varphi(\theta)$ and the matrix of derivatives $\varphi_{\theta'}(\theta)$, and can be implemented numerically if

Table 4: Comparison of the normal approximation to r and to r^*, with r^* using (20), for the Behrens-Fisher problem. Shown are the noncoverages of the nominal upper and lower endpoints for a 90% confidence interval, in 100,000 simulations. In all rows $\mu_1 = 2$, $\mu_2 = 0$.

exact				0.05	0.95	0.05	0.95
n_1	n_2	σ_1^2	σ_2^2	r		r^*	
3	2	2	1	0.015	0.893	0.033	0.965
20	2	2	1	0.127	0.875	0.066	0.934
7	5	2	1	0.069	0.931	0.045	0.950
20	15	2	1	0.057	0.944	0.050	0.950
3	2	4	1	0.104	0.895	0.041	0.959
20	2	4	1	0.109	0.898	0.063	0.938
7	5	4	1	0.058	0.930	0.050	0.949
20	15	4	1	0.957	0.944	0.050	0.950
2	3	2	1	0.124	0.875	0.041	0.959
2	20	2	1	0.155	0.849	0.069	0.933
5	7	2	1	0.074	0.925	0.051	0.949
15	20	2	1	0.057	0.942	0.050	0.949
2	3	4	1	0.137	0.862	0.048	0.951
2	20	4	1	0.162	0.842	0.069	0.933
5	7	4	1	0.077	0.922	0.052	0.948
15	20	4	1	0.058	0.941	0.050	0.949

necessary. This enables approximate inference for parameters in generalised linear models other than linear functions of the canonical parameters. In the two special cases discussed above this simplifies to the versions discussed there; i.e. for inference in generalised linear models with canonical link function Q simplifies to q_E given in (18), and in a regression scale model, the general version reproduces the marginal approximation using q_B (16) (without the prior).

We illustrate with the Behrens-Fisher problem (Cox and Hinkley, 1974, §5.2iv). Suppose we have independent samples of size n_1 and n_2 from normal distributions with parameter $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$, respectively, and the parameter of interest is $\psi = \mu_1 - \mu_2$. The canonical parameter of the full exponential model is

$$\varphi(\theta) = (\mu_1/\sigma_1^2, -1/2\sigma_1^2, \mu_2/\sigma_2^2, -1/2\sigma_2^2).$$

Table 4 gives a comparison of the first- and third- order approximations for a selection of values of n_1, n_2, σ_1^2 and σ_2^2. Except in cases of extreme imbalance, the third order approximation is remarkably accurate.

Given that the calculations can be relatively easily implemented in a variety of models, it would be useful for theoretical discussions to be able to delineate the inferential basis for approximations based on r^* computed using u in (19) or Q in (20), and the situation is still somewhat unsatisfactory. The quantity r^* does seem to arise quite naturally from the likelihood ratio, and the marginal distribution of r^* is what gives a pivotal statistic, but beyond this it is I think difficult to see what 'statistical principle' is at work. Perhaps the asymptotic derivation is enough. Some further discussion of this point is given in recent work by Pierce and Bellio (2004).

The similarity of the approximations in the Bayesian and non-Bayesian versions suggests the possibility of choosing a prior so that the resulting posterior distribution has accurate frequentist coverage to some order of approximation. A discussion of the asymptotic normality of the posterior distribution shows that Bayesian and nonBayesian likelihood inference have the same limiting distributions, so investigating this further requires higher order asymptotics. Welch and Peers (1963) showed for scalar θ that Jeffreys prior $\pi(\theta) \propto \{i(\theta)\}^{1/2}$ ensures that posterior quantiles provide lower confidence bounds with error $O(n^{-3/2})$. Extensions to the nuisance parameter setting have been less successful, as unique priors providing this asymptotic equivalence are not readily available. An overview of these results is given in Reid, Mukerjee and Fraser (2003). An interesting foundational point was raised in the discussion and reply of Pierce and Peters (1992), and is illustrated perhaps most clearly in the derivation of the distribution function for r^* from the p^* approximation to the density of $\hat{\theta}$ given at (7). A change of variable from $\hat{\theta}$ to r means the Jacobian includes

$\partial\ell/\partial\hat{\theta}$, which in turn means that the result involves a very particular sample space dependence of the log-likelihood function.

In the nuisance parameter setting the profile log-likelihood plays an important role, and in particular a plot of the profile log-likelihood can often be informative (Barndorff-Nielsen and Cox, 1994, §3.5). As argued above, adjustments to the profile log-likelihood arise naturally from the asymptotic point of view, although the simple adjustment using $\log|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|$ is unsatisfactory in its dependence on the parameterisation. More refined adjustments to the profile log-likelihood are based on the same asymptotic arguments as lead to the r^* approximation, and a computationally tractable version based on the arguments that lead to (20) is described and illustrated in Fraser (2003).

This discussion has focussed on inference for a scalar component of θ, as that leads to the most tractable, and also most effective, approximations. The p-value function approach does not extend easily to inference about a vector parameter, as departures from the 'null' point are multidimensional. Bartlett correction of the log-likelihood ratio does provide and improvement of the asymptotic χ^2 approximation, which is another motivation for the primacy of the likelihood ratio over asymptotically equivalent statistics such as the standardized maximum likelihood estimate. Skovgaard (2001) proposes a different form of correction of the log-likelihood ratio that is closer in spirit to the r^* approximation. Another possibility is to derive univariate departures in some way, such as conditioning on the direction of the alternative vector and examining its length.

The most important advance in Bayesian inference is the possibility of simulating observations from the posterior density using sophisticated numerical methods. This has driven an explosion in the use of Bayesian methods in practice, particularly for models in which there is a natural hierarchy of effects, some of which can be modelled using priors and hyper-priors. In these settings the inferential basis is straightforward, and the main difficulties are computational, but there is less attention paid to the stability of the inference with respect to the choice of prior than is perhaps desirable. The approximations discussed here can be useful in checking this, and perhaps have a role in assessing convergence of the Markov chain based sampling algorithms. From the point of view of the study of theoretical statistics, an approach which incorporates Bayesian inference as a major component should spend, I think, considerable time on the choice of prior and the effect of the prior. The construction of non-informative priors, and the assessment of the sampling properties of inference based on 'popular' priors should be addressed at length.

3 Discussion

More traditional approaches to the study of the theory of statistics emphasise, to varying degrees, the topics of point estimation and hypothesis testing. The approach based on likelihood essentially replaces both of these by the p-value function. The use of this function also side-steps the discussion of one-tailed and two-tailed tests. The conditional approach to vector parameter inference mentioned above is extends the discussion in Cox and Hinkley (1974, §3.4iv).

Among the more classical topics, some understanding of some basic concepts is needed at least for several specific applied contexts, and perhaps more generally. For example, much of the work in nonparametric regression and density estimation relies on very generally specified models, and relatively *ad hoc* choices of estimators. In this setting the bias and variance of the estimators, usually of functions, but sometimes of parameters, are the only easily identifiable inferential quantities, and it is of interest to compare competing procedures on this basis. As another example, the notion of the power of a test to detect a meaningful substantive difference plays a prominent role in many medical contexts, including the analysis of data from clinical trials. While the relation of power to sample size is less direct, and less meaningful, than the relation of the length of a confidence interval to sample size, the use of fixed level testing is sufficiently embedded in some fields that students do need to learn the basic definitions of size and power. In other fields the basic definitions of decision theory, including loss functions and utilities, will be important.

I don't see any role beyond historical for extended discussion of optimality considerations, particularly with regard to testing. The existence and/or construction of most powerful tests under various special circumstances is in my view largely irrelevant to both practice and theory. The development of admissible or otherwise optimal point estimators may be useful in specialised decision theory settings, and perhaps has a role in choosing among various estimators of functions in settings like nonparametric regression, but does not seem to me to be an essential part of the basics of statistical theory.

One concern about the approach based on higher order asymptotic theory is whether it is excessively specialised, in view of the many developments in modelling for complex settings that have been developed along with the rise in computing power. An important question is whether this theory provides any general concepts that may prove useful

in much more complex settings, such as analysis of censored survival data, of longitudinal data, of data with complex spatial dependencies, of graphical models, or hierarchical models, to name just a few. The approach outlined in §2 emphasises the primacy of the likelihood function, the isolation of a one-dimensional distribution for inference about the likelihood function, and the isolation of a scalar parameter of interest. This can I think be used much more generally.

In many applications of parametric modelling, the role of the likelihood function is well established, and is often the first step in an approach to constructing inference from very complex models. The results described in the previous section indicate that relying on the first-order normal approximation may be misleading, especially in the presence of large numbers of nuisance parameters. Some adjustment for nuisance parameters seems essential, both for plotting the likelihood and for inference. A Bayesian adjustment is the most easily implemented, but raises the difficulty of the dependence of the results on the prior. This dependence is rather clearly isolated in expressions like (14) or (16).

The emphasis in §2 on scalar parameters of interest may be important for discussion of more realistic models, where the identification of this parameter may not be obvious. For example, in the transformed regression model of Box and Cox (1964)

$$y_i^\lambda = x_i'\beta + \sigma e_i, \quad i = 1, \dots n$$

where $\theta = (\beta, \sigma, \lambda)$, and we assume e_i are independent standard normal variates, it seems likely that components of β are not the real parameters of interest. In work as yet unpublished with Fraser and Wong, we suggest that for a simple linear regression model $x_i = (1, z_i)$, one version of a parameter of interest is the rate of change in the median response, at a fixed value z^0:

$$\psi(\theta) = \frac{d}{dz}(\beta_0 + \beta_1 z)^{1/\lambda}|_{z=z^0}.$$

The calculations described in §2.3 are easily carried out in this setting.

In many settings very complex models are constructed at least in part because there is a vast amount of data available. One example is the use of regression models in environmental epidemiology that adjust for potential confounders using methods based on splines. Another is the use of hierarchical modelling of variance components in models for longitudinal data. The study of asymptotic theory can at the least serve as a reminder that the models with very large numbers of nuisance parameters may lead to misleading inference, and that a Bayesian approach is likely to need fairly careful attention to the specification of the prior. The theory can also provide a basis for understanding that the amount of information in the data may be quite different from the apparent sample size. An excellent illustration of this appears in Brazzale (2000, §5.3.1). Recent work by Claeskens (2004) uses marginal or REML likelihood for variance components in fitting regression splines in an interesting combination of ideas from likelihood asymptotics and nonparametric smoothing methods.

In spite of the fact that study of theoretical statistics, and particularly foundations, is perhaps not so fashionable today, there are in our journals a great many theoretical, or at least mathematically technical, papers. Yet most graduate departments seem to be struggling with their basic theory courses, trying to make them relevant, yet finding it difficult to escape the more classical structure. This is certainly an evolutionary process, but I think the time is right to accelerate the modernisation that was initiated in Cox and Hinkley (1974).

Acknowledgements: I would like to thank Augustine Wong for providing the simulation results for the Behrens-Fisher problem and Irene Shi for the simulation results for teh correlation coefficient. I am grateful to Don Fraser and Anthony Davison for helpful discussion, and to two referees for very useful comments.

References:

[1] Barndorff-Nielsen, O.E. (1986). Inference on full and partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322.
[2] Barndorff-Nielsen, O.E., Cox, D.R. (1990). *Asymptotic Techniques for use in Statistics*. Chapman & Hall, London.
[3] Barndorff-Nielsen, O.E., Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.
[4] Barndorff-Nielsen, O.E., Cox, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications (with discussion). *J. R. Statist. Soc. B*, **41**, 279–312.
[5] Bernardo, J.M. and Smith, A.F.M. (2000). *Bayesian Theory*. Wiley, New York.
[6] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc. B*, **26**, 22121–252.

[7] Brazzale, A.R. (2000). *Practical Small-Sample Parametric Inference*. Ph.D. thesis, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland.
[8] Claeskens, G. (2004). Restricted likelihood ratio lack of fit tests using mixed spline models. *J. R. Statist. Soc. B*, **66**, to appear.
[9] Cox, D.R., Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.
[10] Cox, D.R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc. B*, **49**, 1–39.
[11] Cox, D.R., Solomon, P.J. (2002). *Components of Variance*. CRC/Chapman & Hall, Boca Raton.
[12] Davison, A.C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
[13] Davison, A.C. and Wang, S. (2002). Saddlepoint approximations as smoothers. *Biometrika* **89**, 933–938.
[14] Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Math.* **47**, 33–53.
[15] Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65–76.
[16] Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-339.
[17] Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.
[18] Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference from an Neo-Fisherian Perspective*. World Scientific, Singapore.
[19] Pierce, D.A. and Bellio, R. (2004). The effect of choice of reference set on frequency inferences. preprint.
[20] Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. R. Statist. Soc. B*, **54**, 701–737.
[21] Reid, N. (2003).Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695–1731.
[22] Reid, N., Mukerjee, R. and Fraser, D.A.S. (2003). Some aspects of matching priors. *Mathematical Statistics and Applications: Festschrift for C. VanEeden* (M. Moore, S. Froda, C. Léger, eds.) 31–44. Lecture notes Monograph Series 42, Institute of Mathematical Statistics, Hayward.
[23] Schervish, M.J. (1995). *Theory of Statistics*. Springer-Verlag, New York.
[24] Sen, A. and Srivastava, M. (1990). *Regression Analysis Theory, Methods, and Applications*. Springer-Verlag, New York.
[25] Severini, T.A. (2000a). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.
[26] Severini, T.A. (2000b). The likelihood ratio approximation to the conditional distribution of the maximum likelihood estimator in the discrete case. *Biometrika* **87**, 939–945.
[27] Skovgaard, I.M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–165.
[28] Skovgaard, I.M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28**, 3–32.
[29] Welch, B. and Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc. B* **25**, 318–329.