Statistical Science
1995, Vol. 10, No. 2, 138–199

The Roles of Conditioning in Inference

N. Reid

Abstract. This paper reviews the use of sufficient and ancillary statistics in constructing conditional distributions for inference about a parameter. Special emphasis is given to recent developments in accurate approximation of densities, distribution functions and likelihood functions, and to the role of conditioning in these approximations. Exact conditional or marginal inference is available for essentially two classes of models, exponential family models and transformation family models. The approximations are very useful for practical implementation of these exact results. The form of the approximations suggests methods for inference in more general families.

Key words and phrases: Ancillary, exponential models, likelihood, modified likelihood, nuisance parameters, profile likelihood, tail area approximations, transformation models.

1. INTRODUCTION

Most statisticians, theoretical and applied, use conditioning arguments every day in their work without a second thought. Regression analyses in observational studies are amost always performed conditionally on the observed values of the covariates. With some few exceptions all analyses condition on the observed sample size. Conditional probability is such a natural way of separating what we know from what we want to know, that it is hard to imagine statistical science advancing far without it.

"Conditional inference," the topic, though, has a reputation as a vaguely determined area of theoretical statistics, requiring familiarity with such murky concepts as "relevant subsets," "approximate ancillarity," "similar regions" and the like: an area to be avoided if at all possible, and one that is given short shrift in most statistics textbooks. Discussion of conditional inference usually begins with a relatively straightforward example, such as Cox's (1958a) example of two weighing machines of unequal precision. Unfortunately the next and succeeding examples are much less clearcut. The most familiar of these is perhaps Fisher's exact test for 2×2 tables; we seem to have been arguing about it for more than 50 years.

Nancy Reid is Professor, Department of Statistics, University of Toronto, Toronto, Ontario M5S 1A1, Canada.

The distinction between conditioning as a very natural tool in applications and the study of "conditional inference" is, I think, both unfortunate and unnecessary. Conditioning is as natural a tool in the theory of inference as it is in practice. Although there are valid foundational arguments leading to conditioning in particular classes of problems, a very important aspect of conditioning often overlooked in theoretical treatments is that conditional distributions can be much easier to calculate than marginal distributions, as they do not require high-dimensional integration. This fact is powerfully exploited in recent work in spatial statistics and in the use of Markov chain Monte Carlo methods in Bayesian inference.

In the past 15 years there has been rapid development in the area of higher-order asymptotics, which I take here to describe the derivation of more accurate distributional approximations for likelihood-based inference. These approximations are derived from asymptotic expansions which include terms beyond the leading term. Conditioning is an essential feature of this work. Of course it would not usually be of interest to construct a conditional model which could be accurately approximated if such a model did not provide "good" inference for θ, in some sense. A major feature of recent results in likelihood-based inference is that conditional densities which are suitable for inference about θ can be accurately approximated.

Cox (1988) identifies at least four interrelated roles for conditioning in inference: to make probability calculations relevant to the data under

study, to recover information lost in reducing the dimension of the problem, to eliminate nuisance parameters and to enable computation of accurate approximations to densities. It is the last of these that has seen very rapid development recently, but it is the first three that enable one to say the approximations are useful for inference. Making probability calculations relevant and recovering information are usually associated with conditioning on ancillary or approximately ancillary statistics; eliminating nuisance parameters is usually associated with conditioning on sufficient or approximately sufficient statistics. These two types of conditioning are most transparent in transformation family models and exponential family models, respectively.

In the remainder of this section, we consider in detail the notions of sufficiency and ancillarity, in the simple case of no nuisance parameters. In Section 2 we summarize approximate calculations for this setting. Section 3 discusses sufficiency and ancillarity in the presence of nuisance parameters, and Section 4 discusses accurate approximation for models with nuisance parameters. Section 5 reviews extensions of exact sufficiency and ancillarity to approximate sufficiency and ancillarity in more general settings. Section 6 examines some other uses of conditioning in inference and touches on some foundational aspects such as the relation between conditioning and Bayesian inference and the relation between conditioning and power.

Throughout we consider parametric models for a random vector y and inference for the parameter θ or for some component of the parameter θ. To fix some ideas and notation, we first consider the case where inference is to be constructed for the parameter θ. Recall that if there exists a one-to-one transformation from y to (s, t) such that

$$f(y;\theta) \propto f(s;\theta)f(t|s), \tag{1.1}$$

then $s = s(y)$ is sufficient for θ and inference for θ is based on the marginal density for s. In (1.1) the Jacobian of the transformation from y to (s, t) is absorbed into the proportionality constant.

The conditional density in (1.1) does have a role to play in inference, but not in inference for θ. As is suggested in Cox and Hinkley (1974, Chapter 2) the conditional density $f(t|s)$ is useful for model checking; in particular because it does not depend on which value of θ generated the data y. From the model-checking point of view, θ is a nuisance parameter which is eliminated in the conditional density.

The most well known class of models which allows a sufficiency reduction of the type (1.1) is the family of linear exponential models $f(y;\theta) = \exp\{\theta' s(y) - k(\theta) - d(y)\}$; $s = s(y)$ is the minimal sufficient statistic and is of the same dimension as the parameter θ.

In contrast to (1.1), if

$$f(y;\theta) \propto f(s|t;\theta)f(t), \tag{1.2}$$

then $t = t(y)$ is said to be ancillary for θ. In this case it is the conditional density that is used for inference about θ. The argument often advanced for using $f(s|t;\theta)$ for inference about θ is either that the conditional density gives more precise inference for θ or that the subset of the sample space defined by fixing the value of t is the relevant subset for inference about θ. These were the arguments advanced in Fisher (1934); for further discussion see Dawid (1991) and Kass (1989). The idea of using the conditional distribution in (1.2) for inference has not been as widely accepted as using the marginal density in (1.1). This may be because the partition in (1.2) is not typically unique, whereas that in (1.1) is essentially unique.

The main class of models which allows an ancillary reduction of the type (1.2) is the class of transformation models, that is, models generated by a group of transformations $\mathscr{G}$, say, acting on the sample space. Denoting the base model by $f_0(\cdot)$, the transformation model is $f(y;\theta) = f_0(g_\theta y)$, where g_θ is a member of a group of transformations indexed by θ that leaves the original sample space unchanged. For example, in a location–scale model, $g_\theta y = (y - \theta_1)/\theta_2$. In a sample of size n from a transformation model a partition of the form (1.2) obtains with $s = s(y) = \hat{\theta}$, the maximum likelihood estimate, and $t = t(y) = g_{\hat{\theta}} y$, the maximal invariant for the group. Again s has the same dimension as the parameter θ. In a sample of size n from a location model, $t(y) = (y_1 - \hat{\theta}, \ldots, y_n - \hat{\theta})$; for a location–scale model, $t(y) = (\hat{\theta}_2^{-1}(y_1 - \hat{\theta}_1), \ldots, \hat{\theta}_2^{-1}(y_n - \hat{\theta}_1))$.

In transformation family models, where (1.2) obtains, the conditional density $f(s|t;\theta)$ can be obtained by renormalizing the likelihood function (Fisher, 1934; Fraser, 1968, Chapter 2; Fraser, 1979, Chapter 7; Efron and Hinkley, 1978; Barndorff-Nielsen, 1980). In exponential family models where (1.1) obtains, the marginal density $f(s;\theta)$ can be obtained, at least to a high degree of approximation, by renormalizing the likelihood function (Daniels, 1954, 1958; Barndorff-Nielsen and Cox, 1979). In 1980 a series of papers in *Biometrika* (Cox, 1980; Barndorff-Nielsen, 1980; Durbin 1980a, b; Hinkley, 1980a) discussed aspects of approximate sufficiency and ancillarity and the approximation of the relevant distributions by the renormalized likelihood function. A common feature in all this work was a formula that has come to be known as

Barndorff-Nielsen's formula, or the p^* formula. The p^* formula is discussed in the following section.

2. COMPUTATION OF MARGINAL AND CONDITIONAL DISTRIBUTIONS

2.1 The p^* Formula

Barndorff-Nielsen's formula, or the p^* formula, provides an approximation to the conditional density of the maximum likelihood estimate $\hat{\theta}$, given an auxiliary statistic t:

$$(2.1)\quad f(\hat{\theta}|t;\theta) \doteq c\left\{\frac{L(\theta;\hat{\theta},t)}{L(\hat{\theta};\hat{\theta},t)}\right\}|j(\hat{\theta})|^{1/2} \equiv p^*(\hat{\theta}|t).$$

In (2.1) it is important to note that the likelihood function $L(\theta;\hat{\theta},t)$ has been expressed as a function of $(\hat{\theta},t)$ in its dependence on the data; that is, using the usual definition of the likelihood function as proportional to the sampling density,

$$L(\theta;y) = a(y)f(y;\theta),$$

it has been assumed that the minimal sufficient statistic can be expressed as a one-to-one function of $\hat{\theta}$ and t. The observed Fisher information appearing in (2.1) is defined by

$$j(\hat{\theta}) = -\left\{\frac{\partial^2 \log L(\theta;y)}{\partial\theta\,\partial\theta^T}\right\}\Bigg|_{\theta=\hat{\theta}}$$
$$= -\left\{\frac{\partial^2 \log L(\theta;\hat{\theta},t)}{\partial\theta\,\partial\theta^T}\right\}\Bigg|_{\theta=\hat{\theta}}.$$

In discussion of the p^* formula it is usually assumed that the marginal distribution of t is free of θ, at least approximately; that is, that t is an exactly or approximately ancillary statistic.

It is essentially (2.1) that highlights the role of conditioning in constructing accurate approximations (Barndorff-Nielsen, 1988b, Chapter 1). The requirement that t be approximately ancillary ensures that there is little information about θ in the marginal density for t, so that the conditional distribution of $\hat{\theta}$ given t is expected to provide "good" inference for θ. The existence of an accurate approximation to this distribution enables computation of significance probabilities, confidence intervals and so on.

In the special case that $f(y;\theta)$ is a transformation model, a maximal ancillary exists and factorization (1.2) holds. In this case, with t fixed, $\hat{\theta}$ is a one-to-one function of s, and the exact conditional distribution of $\hat{\theta}$ given t is given by (2.1) (Fraser, 1968, Chapter 2; Barndorff-Nielsen, 1980). As a simple example, in a sample of size n from the location model, the exact conditional density is

$$(2.2)\quad f(\hat{\theta}|t) = c\prod_{i=1}^{n}\left\{\frac{f_0(t_i+\hat{\theta}-\theta)}{f_0(t_i)}\right\},$$

where c is a normalizing constant for the conditional density. This result is given in Fisher (1934); see also Cox and Hinkley (1974, Chapter 7).

In the special case that $f(y;\theta)$ is an exponential family model, the sufficient statistic $s(y)$ is a one-to-one function of $\hat{\theta}$; thus the likelihood function depends on the data only through $\hat{\theta}$. In this case (2.1) gives an approximation to $f(\hat{\theta};\theta)$ with relative error $O(n^{-3/2})$ and can be derived directly from the saddlepoint approximation to the density of the sample mean (Daniels, 1954, 1958; Barndorff-Nielsen and Cox, 1979; Barndorff-Nielsen, 1983). These results are reviewed in more detail in Reid (1988).

Formula (2.1) holds quite generally, subject to specification of an approximate ancillary. A detailed discussion and review of the literature is given in Barndorff-Nielsen and Cox (1994, Chapter 6), and a proof of its validity in curved exponential families is given in Barndorff-Nielsen and Cox (1994, Section 7.4). An elegant and very general proof is given in Skovgaard (1990), by first deriving an expression for the exact distribution of the maximum likelihood estimator as a product of three terms; one is the likelihood ratio that appears in (2.1); two others combine to give the information determinant and the normalizing constant, after approximation (see, in particular, Skovgaard, 1990, equation (3.3)).

2.2 Tail Area Approximations

In the case that θ is a scalar, there are also available accurate approximations to the cumulative distribution function for $\hat{\theta}$, which may be more useful in applications. The general tail area approximation is

$$(2.3)\quad F(\hat{\theta}|t;\theta) = \left\{\Phi(r)+\varphi(r)\left(\frac{1}{r}-\frac{1}{u}\right)\right\}\{1+O(n^{-3/2})\},$$

where

$$(2.4)\quad r = \pm\left(2\left[\log\left\{\frac{L(\hat{\theta})}{L(\theta)}\right\}\right]\right)^{1/2},$$

$$(2.5)\quad u = \left\{\frac{\partial l(\theta;\hat{\theta},t)}{\partial\hat{\theta}}\Bigg|_{\theta=\hat{\theta}} - \frac{\partial l(\theta;\hat{\theta},t)}{\partial\hat{\theta}}\right\}|j(\hat{\theta})|^{-1/2},$$

$l(\theta) = \log L(\theta)$ and Φ and φ are the standard normal distribution and density functions, respectively. This result is obtained in Barndorff-Nielsen (1988a, 1990) and Fraser (1990), from the p^* formula (2.1). In order to use the approximation it is necessary

to compute the derivative in (2.5) with the approximate ancillary statistic t from (2.1) held fixed.

Again the approximation simplifies in transformation and exponential families. In transformation families, such as the location model, t is the maximal invariant and u simplifies to the standardized score statistic:

$$u = v = l'(\theta)|j(\hat{\theta})|^{-1/2}. \tag{2.6}$$

In exponential families the maximum likelihood estimate is sufficient, no ancillary statistic is needed and u simplifies to the Wald statistic:

$$u = q = (\hat{\theta} - \theta)|j(\hat{\theta})|^{1/2}. \tag{2.7}$$

The tail area approximation (2.3) using $u = q$ was derived in Lugannani and Rice (1980) from the saddlepoint approximation to the density of $\hat{\theta}$; see also Daniels (1987).

An alternative version of (2.3) can be obtained by approximating the quantile. Let

$$r^* = r + \frac{1}{r}\log\frac{u}{r}. \tag{2.8}$$

Then

$$F(\hat{\theta}|t;\theta) = \Phi(r^*)\{1 + O(n^{-3/2})\}; \tag{2.9}$$

this can be verified using a Taylor expansion of r about u (Barndorff-Nielsen, 1990; Jensen, 1992). The r^* version, due to Barndorff-Nielsen (1986) can sometimes be more accurate than (2.3) although the difference is often slight.

2.3 Approximate Ancillarity

To use either the p^* formula or the tail area formula (2.3) or (2.8), it is necessary to be able to find an approximately ancillary statistic t such that $(\hat{\theta}, t)$ is a one-to-one function of the minimal sufficient statistic y. In many practical problems this will be either difficult or impossible, although in special settings a variety of approximately ancillary statistics have been suggested.

"Approximately ancillary" is usually taken to mean that the density of t does not depend on θ, for θ in an $\sqrt{n}$-neighborhood of the true value. Following Cox (1980) we say that t is qth-order locally ancillary if

$$f(t;\theta_0 + n^{-1/2}\delta) = f(t;\theta_0)\{1 + O(n^{-q/2})\}.$$

A fairly general construction of second-order locally ancillary statistics in the scalar parameter case is given in McCullagh (1987, Section 8.3), following the development in Cox (1980) and McCullagh (1984). There is no unique second-order locally ancillary statistic, but the p^* formula can be shown to hold conditionally on any of the second-order ancillary statistics, with relative error $O(n^{-1})$ (cf. McCullagh, 1987, Section 8.6, where the same result is obtained to $O(n^{-3/2})$ using third-order ancillary statistics). McCullagh (1984) uses the term locally sufficient for a statistic which is independent of a local ancillary and shows that the signed square root of the likelihood ratio statistic can be adjusted by a bias correction to be locally sufficient to order $O(n^{-1})$. Skovgaard (1990) gives a very thorough and helpful discussion of approximate ancillarity in relation to the p^* formula and indicates that the p^* formula will hold with relative error $O(n^{-3/2})$ when the statistic t is ancillary only to second order.

The construction of locally ancillary statistics outlined in McCullagh (1987, Section 8.3) and Cox (1980) is based on linear combinations of log-likelihood derivatives, evaluated at θ_0. A similar construction of locally ancillary statistics based on linear combinations of log-likelihood derivatives evaluated at $\hat{\theta}$ is outlined in detail in Barndorff-Nielsen and Cox (1994, Section 7.2). For both constructions the main idea is that the array of log-likelihood derivatives $(l_\theta, l_{\theta\theta}, l_{\theta\theta\theta}, \ldots)$ is jointly asymptotically normal, can be standardized to be jointly asymptotically standard normal and thus to have a distribution not depending on θ, at least approximately. The order of ancillarity is determined by the order of the highest log-likelihood derivatives used in constructing the statistic. The standardization involves computation of the joint cumulants of the likelihood derivatives. For example, McCullagh (1987, equations (8.7) and (8.8)) derives the linear combination of $l_\theta(\theta_0)$ and $l_{\theta\theta}(\theta_0)$ that is ancillary to second order. The function of $\hat{\theta}$ and $l_{\theta\theta}(\hat{\theta})$ that is ancillary to second order is the Efron–Hinkley ancillary $(\hat{\imath}\hat{\gamma})^{-1}(\hat{\jmath} - \hat{\imath})$, where i and j are the per-observation expected and observed information, respectively, and $\hat{\gamma} = \gamma(\hat{\theta})$ is an estimate of the Efron or exponential curvature of the model (Barndorff-Nielsen and Cox, 1994, equation (7.1)). In a scalar parameter family γ is the residual variance of $l_{\theta\theta}$, after linear regression on l_θ (Hinkley, 1980a).

When the model of interest can be viewed as a submodel of an exponential or transformation model, possibly by introducing nuisance parameters, it is usually possible to obtain an approximately ancillary statistic from the larger model. This is the approach taken, for example, in Barndorff-Nielsen (1984) in constructing approximately ancillary statistics for curved exponential families. This same construction can be used to eliminate nuisance parameters by approximate ancillarity, as is done in Barndorff-Nielsen (1986).

Two types of approximate ancillaries using this approach are derived in Barndorff-Nielsen and Cox (1994, Section 7.2); the first based on linear combinations of log-likelihood derivatives, and the second based on the signed square root of the likelihood ratio statistic for testing the model of interest against the larger model.

EXAMPLE 2.1. This example is discussed in Barndorff-Nielsen and Cox (1994, Exercise 7.4), and in Hinkley (1980a) and Skovgaard (1985). Let (Y_i, Z_i), $i = 1, \ldots, n$, be independent, identically distributed with $Z_i \sim N(\theta, 1)$ and $Y_i|Z_i \sim N(\theta_2 Z_i, 1)$. The log-likelihood function is

$$l(\theta_1, \theta_2) = -\tfrac{1}{2}\Big\{(\theta_2^2+1)\textstyle\sum z_i^2 - 2\theta_1 \sum z_i - 2\theta_2 \sum y_i z_i + n\theta_1^2\Big\} \tag{2.10}$$

with minimal sufficient statistic $(\sum y_i z_i, \sum z_i^2, \sum z_i)$. As indicated in Barndorff-Nielsen and Cox (1994), the Efron–Hinkley ancillary is equivalent to the exact ancillary, which is a mean and variance standardization of $a = \sum(z_i - \bar{z})^2$. The ancillary constructed in McCullagh (1987, Section 8.3) is obtained from a linear combination of $\partial l/\partial\theta_1$, $\partial l/\partial\theta_2$ and $V_{22} = \partial^2 l/\partial\theta_2^2 + 2\theta_1(\partial l/\partial\theta_1) - 2\theta_2(\partial l/\partial\theta_2) = (2\theta_2^2 - 1)\sum z_i^2 - 2\theta_2 \sum z_i y_i + 2\theta_1 \sum z_i$. The resulting statistic is locally ancillary at θ_0 when evaluated there, although it seems most natural to evaluate it at $\hat{\theta}$, in which case it simplifies to a statistic equivalent to the exact ancillary.

Note that the log-likelihood function is that of a (3, 2) curved exponential family, although it does not seem obvious what the most natural embedding (3, 3) family is. For example, we could simply introduce a third parameter θ_3 as the coefficient of $\sum z_i^2$. It is then necessary to compute the normalizing constant for the joint density in order to evaluate the likelihood ratio statistic for testing this larger model against (2.10). While this is not too difficult in this example, it leads to rather unwieldy expressions.

If we are interested in using the r^* or tail area approximation, rather than the p^* formula, then it is not necessary to determine t explicitly. It is only necessary to be able to differentiate the log-likelihood function with t held fixed, that is, to determine the gradient of the likelihood in approximately ancillary directions. A general technique for computing this is outlined in Fraser and Reid (1995). For scalar parameters, the ancillary statistic is simply $\{-\partial F(y_i;\theta)/\partial\theta\}/\{\partial F(y_i;\theta)/\partial y_i\}$, evaluated at $\theta = \hat{\theta}$, where F is the cumulative distribution function. The vector parameter case is more difficult and is obtained only for a class of generalized linear models. In addition, it is shown in Fraser and Reid (1995) that for computing tail area approximations accurate to $O(n^{-3/2})$ it is sufficient to obtain the gradient to $O(n^{-1})$.

The development of Barndorff-Nielsen's formula and techniques for constructing approximate ancillaries provide an accurate means for implementing conditional inference and, in the setting discussed in this section, of inference for the whole parameter θ, there is little or no controversy in implementing this conditional approach. However, in constructing inference for components of θ in the presence of nuisance parameters, the controversial examples, such as inference for the odds ratio in a 2×2 table, are examples involving nuisance parameters. While progress has been made in accurate approximation for tail probabilities in the presence of nuisance parameters, as we shall see in the next section the role of conditioning is much less clear than it is in (1.1) and (1.2).

3. NUISANCE PARAMETERS

We will assume that the parameter θ has two components $[\theta = (\psi, \lambda)]$ and that inference is to be constructed for one component (usually ψ) with the other component treated as a nuisance parameter. Nuisance parameters arise in a variety of settings, but typically they are included to make the model more realistic for the application of interest.

We assume that if the model $f(y;\theta)$ satisfies (1.1) or (1.2), the marginal or conditional density for s will be used as the basis for inference, and we will no longer explicitly indicate the conditioning in (1.2). In other words, we have a reduction in dimension from the original problem to the dimension of s. In exponential or transformation models, the dimensions of s and θ are the same.

In considering extensions of factorizations (1.1) and (1.2), the simplest is

$$f(s;\psi,\lambda) = f(s_1|s_2;\psi)f(s_2;\lambda). \tag{3.1}$$

Even in this simple setting, the definitions of sufficiency and ancillarity are not unambiguous. For example, we could say that s_2 is sufficient for λ, as in (1.1), or that s_2 is ancillary for ψ, as in (1.2).

Many definitions of sufficiency and ancillarity in the presence of nuisance parameters require the parameters of the model to split into component densities in the manner of (3.1) (Fraser, 1956; Basu, 1977; Cox and Hinkley, 1974, Chapter 2). Cox and Hinkley (1974, Chapter 2) refer to s_1 as "conditionally sufficient for ψ." Barndorff-Nielsen (1978, Chapter 4)

uses the terminology "s_1 is S-sufficient for ψ." Basu (1977, 1978) calls s_2 partially sufficient for λ.

Whatever difficulties with definitions might exist, it is clear from (3.1) that we should use the conditional distribution of s_1 given s_2 for inference about ψ, and the marginal distribution of s_2 for inference about λ. In the first case the nuisance parameter λ is eliminated by conditioning, and in the second the nuisance parameter ψ is eliminated by marginalizing.

EXAMPLE 3.1. Suppose that Y_1 and Y_2 are independent Poisson random variables with means φ and $\varphi\psi$, respectively. The conditional distribution of Y_1 given $Y_+ = Y_1 + Y_2$ is binomial with probability of success $\psi/(1+\psi)$, and the marginal distribution of Y_+ is Poisson with parameter $\varphi + \varphi\psi = \lambda$, say. So (3.1) holds:

$$f(y_1, y_2; \theta) = f(y_1|y_+; \psi) f(y_+; \lambda);$$

inference for ψ is constructed from the conditional distribution of Y_1 given the total Y_+; and inference for λ is constructed from the marginal distribution of Y_+. If we think of Y_1 and Y_2 as the counts in two Poisson processes, it is intuitively obvious that there is no information available on the ratio of the rates from the total count $Y_1 + Y_2$.

Note that the parameter λ was redefined from the original parametrization in order for (3.1) to apply. It is essential in this redefinition that the new parametrization includes all of the original parameter space, or that ψ and λ be variation independent. In this situation the statistic s_2 in (3.1) is called a cut (Barndorff-Nielsen, 1978, Chapter 4).

Very few models with nuisance parameters admit a factorization of the form (3.1). One generalization of (3.1) is

$$f(s; \theta) = f(s_1|s_2; \psi) f(s_2; \psi, \lambda). \tag{3.2}$$

Now s_2 is no longer ancillary for ψ, but it is sufficient for λ in the sense that the nuisance parameter λ has been eliminated in the conditional distribution. One motivation for using $f(s_1|s_2; \psi)$ for inference about ψ is merely pragmatic: we can do this without specifying a value for the unknown parameter λ. A more theoretical motivation is that in testing an hypothesis about ψ, with λ unspecified, any test having a type I error that does not depend on λ must be constructed from the conditional distribution, at least if s_2 is complete (Lehmann, 1986, Chapter 4). However, there is potentially information about ψ in the marginal density of s_2, as is indicated explicitly in the notation. A systematic investigation into ways of quantifying the information in such marginal (or conditional) densities is given in Barndorff-Nielsen (1978, Chapter 4) and in recent work by Jorgensen (1993).

EXAMPLE 3.2. The most common models admitting a factorization of the form (3.2) are exponential family linear models

$$f(s; \theta) = \exp\{\psi s_1 + \lambda s_2 - k(\psi, \lambda) - d(s_1, s_2)\}.$$

It is easy to show that

$$f(s_1|s_2; \psi) = \exp\{\psi s_1 - k_2(\psi) - d_2(s_1)\} \tag{3.3}$$

and that the marginal distribution of s_2 depends on (ψ, λ) (Lehmann, 1986, Chapter 2). The Poisson model of example (3.1) is a special case of this. In (3.3) the functions $k_2(\psi)$ and $d_2(s_1)$ depend on s_2 and are usually difficult to calculate. From Section 2 we know that calculation of $d(s_1, s_2)$ can be bypassed by the saddlepoint approximation and would thus expect that calculation of $d_2(s_1)$ can be similarly bypassed. In fact, as we shall see in Section 4, calculation of $k_2(\psi)$ can also be bypassed this way. For applications it will be important to generalize to the case where the parameter of interest is not a component of the canonical parameter, and this is discussed briefly in Section 5.

Tests of hypotheses about ψ based on (3.3) have an unconditional optimality property: they are uniformly most powerful among the class of unbiased tests. Conditional inference based on (3.3) is discussed in detail in Lehmann (1986, Section 4.4) from this point of view. Examples 3.1 and 3.4 are considered in Section 4.5, and Example 3.3 is given as a problem in Section 5.16.

EXAMPLE 3.3. The shape parameter of a gamma distribution is a component of the canonical parameter. Suppose we have a sample of size n from the density

$$f(y; \psi, \lambda) = \frac{1}{\Gamma(\psi)} \lambda^{\psi} y^{\psi - 1} \exp(-\lambda y). \tag{3.4}$$

The minimal sufficient statistic is

$$(s_1, s_2) = \left(\sum \log y_i, \sum y_i\right)$$

and the conditional density of s_1 given s_2 is given by (3.3), where versions of $k_2(\cdot)$ and $d_2(\cdot)$ are

$$\begin{aligned} \exp\{k_2(\psi)\} &= \int \exp\{(\psi - 1)s_1\} h(s_1, s_2)\, ds_1, \\ \exp\{-d_2(s_1)\} &= \exp(-s_1) h(s_1, s_2), \\ h(s_1, s_2) &= \exp\{-d(s_1, s_2)\} \\ &= \int_A dy_1 \cdots dy_n, \end{aligned} \tag{3.5}$$

where $A = \{(y_1, \ldots, y_n)\colon \sum y_i = s_2, \sum \log y_i = s_1\}$.

EXAMPLE 3.4. Another special case of Example 3.2 is the comparison of two binomials. Let $Y_1 \sim \text{Bin}(n_1, p_1)$ and $Y_2 \sim \text{Bin}(n_2, p_2)$ be independent, and suppose that the parameter of interest is the log odds ratio $\psi = \log[p_1(1-p_2)/\{p_2(1-p_1)\}]$. This model is an exponential linear model with $\lambda = \log\{p_2/(1-p_2)\}$. Then, writing $Y_+ = Y_1 + Y_2$,

$$f(y_1, y_2; \theta) = f(y_1|y_+; \psi) f(y_+; \psi, \lambda),$$

where the first factor is the noncentral hypergeometric density

$$f(y_1|y_+; \psi) = \binom{n_1}{y_1}\binom{n_2}{y_+ - y_1} \exp(\psi y_1)/C(\psi, y_+).$$

The normalizing constant $C(\psi, y_+)$ is the function $\exp\{k_2(\psi)\}$ of (3.3) and is typically cumbersome to compute exactly. Specialized programs are now available for exact computation of $C(\psi, y_+)$, however, and are reviewed in Agresti (1992). Extension of this example to logistic regression is considered in Cox (1958b) and Cox and Hinkley (1974, Chapter 5).

Use of $f(y_1|y_+; \psi)$ when $\psi = 0$ corresponds to Fisher's exact test for the equality of two binomial proportions, or for the absence of interaction in a 2×2 table, and has a long history of controversy associated with it. A recent reference that touches on many of the issues is Yates (1984), but see also the several references in Agresti (1992, Section 8.1). If we array the data in the form of a 2×2 table, the four cell entries are y_1, y_2, $n_1 - y_1$ and $n_2 - y_2$, and conditioning on y_+ means ignoring information about ψ in the margins of the table. That there is information about ψ in the distribution of y_+ is clear, but quantifying it has proved rather elusive (Plackett, 1977; Yates, 1984). The partial Fisher information for ψ in the marginal distribution of y_+ is computed in Zhu and Reid (1994).

Generalizations of the 2×2 table to loglinear models for multidimensional contingency tables also admit a factorization of the form (3.1), with s_2 the vector of marginal totals. Conditional inference for these models is reviewed in Agresti (1992), and several discussants indicate unease in basing inference on $f(s_1|s_2; \psi)$. As emphasized by Agresti in his reply to the discussion, and in Barnard (1984), the issue is clouded by the discreteness of the conditional simple space, and it seems likely that careful investigation of the properties of mid-p values will help to resolve this aspect of the controversy.

EXAMPLE 3.5. An important extension of Example 3.4 is to consider a series of independent pairs of binomials, each with the same value of ψ but different values of λ, and each with small sample sizes, say $n_1 = n_2 = 1$. The maximum likelihood estimate of ψ based on the joint density of $(y_{1i}, y_{2i}; i = 1, \ldots, k)$ is inconsistent as $k \to \infty$; in fact $\hat{\psi}$ converges to 2ψ as $k \to \infty$ (Andersen, 1973; Breslow, 1981; Breslow and Day, 1980, Chapter 7; Jennison, 1992). Use of the conditional density $\Pi f(y_{1i}|y_{+i}; \psi)$ provides a consistent estimate of ψ. However, this estimate of ψ can be very inefficient if in fact the λ_i are all equal (Liang, 1987). Evidence for the equality of the λ_i or subsets of them might be available extraneously and incorporated as a prior, or it might be obtained by estimating the λ_i's using the marginal density of y_{+i}.

This example is one of a class of problems, sometimes called Neyman–Scott problems, where the dimension of the nuisance parameter increases with the number of observations (or more generally with the Fisher information). One reasonably general version is to assume that we have observations $(y_{ji};\ j = 1, \ldots, n_i;\ i = 1, \ldots, k)$ with $f(y_{ji}; \psi, \lambda_i)$ the model for observations in group i. As $k \to \infty$, the asymptotic theory of Sections 2 and 4 does not apply, and as yet there is no general asymptotic treatment of these models beyond the first-order theory outlined in Portnoy (1988). Note that although the dimension of the nuisance parameter is unbounded, it is reasonable to expect that an asymptotic theory could be developed for the case that n_i increases with k. Some examples are discussed in Cox and Reid (1992).

A different generalization of (3.1) is the case where s_2 is no longer sufficient for λ, but is ancillary for ψ:

$$f(s; \psi, \lambda) = f(s_1|s_2; \psi, \lambda) f(s_2; \lambda). \tag{3.6}$$

By analogy with the situation in (3.2), for inference about λ, we would use the marginal density of s_2. Again, one motivation for using this marginal density is pragmatism; we can construct inference for λ without specifying any value for the nuisance parameter ψ. If our interest is in the parameter ψ, then we will either use $f(s; \psi, \lambda)$ or $f(s_1|s_2; \psi, \lambda)$ for inference about ψ. If we use the conditional density, plausible values of λ might be obtained from the marginal density for s_2. This is a more direct generalization of the ancillarity definition (1.2): the distribution of the ancillary statistic depends on an additional parameter rather than being completely known as in (1.2).

EXAMPLE 3.6. Suppose we have a sample of size n from the normal distribution with mean μ and variance σ^2. Writing $\bar{y}$ and s^2 for the sample mean

and variance, we have

$$f(\bar{y}, s^2; \mu, \sigma^2) = f_1(\bar{y}; \mu, \sigma^2) f_2(s^2; \sigma^2),$$

where f_1 is the $N(\mu, \sigma^2/n)$ density and f_2 is the $\sigma^2 \chi^2_{n-1}$ density. Because $\bar{y}$ and s^2 are independent, this is a generalization of both (3.2) and (3.6). Inference about σ^2 is based on the marginal distribution of s^2 [as in (3.6), s^2 is ancillary for μ] or the conditional distribution of s^2 given $\bar{y}$ [as in (3.2), $\bar{y}$ is sufficient for μ]. In fact this example highlights the difficulty with the extended definitions of sufficiency and ancillarity. Although $\bar{y}$ is sufficient for μ in the sense of definition (3.2), it is not sufficient in our usual understanding of the definition based on (1.1); that is, inference for μ cannot be constructed from $\bar{y}$ alone.

Of course in this example inference for μ is constructed from the t-statistic $\sqrt{n}(\bar{y} - \mu)/s$. The t-test can be derived by formal considerations related to (3.2) and (3.6). One derivation is via the construction of similar tests for the ratio of canonical parameters in the exponential family (discussed briefly in Section 5; see also Lehmann, 1986, Section 5.2) and the other is via construction of an invariant test (Lehmann, 1986, Section 6.4). The t-statistic is the most well-known example of a pivotal statistic, that is, a function of the data and parameters whose distribution is known exactly. The development of inference based on pivotal statistics proceeded somewhat separately from that of conditional inference, although recent work emphasizes the connections between them. The normal distribution is both an exponential family and a transformation family, which is why arguments based on sufficiency or ancillarity lead to the same result.

In transformation models tests based on the marginal distribution of the ancillary statistic (the maximal invariant) have an unconditional optimality property: they are uniformly most powerful among the class of invariant tests. This is the point of view from which the t-test is derived in Lehmann (1986, Section 6.4).

EXAMPLE 3.7. As a slight generalization of the location family, suppose we have a sample of n observations following the regression model

$$Y = X\beta + \varepsilon, \tag{3.7}$$

where X is an $n \times p$ matrix of regression constants, and ε has a known density $f_0(\varepsilon)$. As in the location model, an ancillary statistic is available and a preliminary reduction by conditioning is assumed. The (conditional) density $f(\hat{\beta}; \beta)$ is itself a location family, that is, $f(\hat{\beta}; \beta) = f(\hat{\beta} - \beta)$. If one component of β is of particular interest, with the remaining components treated as nuisance parameters, then (3.6) obtains:

$$f(\hat{\beta} - \beta) = f(\hat{\beta}_1 - \beta_1) f(\hat{\beta}_{(2)} - \beta_{(2)} | \hat{\beta}_1 - \beta_1), \tag{3.8}$$

where $\beta_{(2)}$ is the nuisance parameter $(\beta_2, \ldots, \beta_p)$. Inference for β_1 is based on the marginal density of $\hat{\beta}_1$, and any potential information about β_1 in the conditional density is ignored. Generalizations of (3.7) that allow incorporation of scale and shape parameters in the density of ε are discussed in Section 5. Conditional inference in linear regression models is discussed in Fraser (1979, Chapter 4) and Lawless (1982, Chapter 6).

Although there is potentially information about β_1 in the conditional distribution in (3.6), this does not seem to have generated any controversy in the literature about the appropriateness of using the marginal distribution. Marginal distributions of the form illustrated in (3.6) also play an important role in Bayesian inference in the presence of nuisance parameters. In the Bayesian setting nuisance parameters are integrated out of the joint posterior density. Recent advances in evaluating high-dimensional integrals have made using marginal densities for inference possible for applications involving large numbers of nuisance parameters.

4. APPROXIMATIONS IN MODELS WITH NUISANCE PARAMETERS

4.1 Introduction

In Section 2 we saw that accurate approximations for conditional or marginal densities and distribution functions can be readily constructed from the likelihood function. These have recently been generalized to the nuisance parameter setting, and the existence of accurate approximations in these settings provides one of the motivations for the discussion in Section 3. In Section 2 we also saw that the approximation for the marginal distribution of a sufficient statistic and the approximation for the conditional distribution given an ancillary statistic had essentially the same form: this is also the case here.

4.2 Exponential Linear Models

As mentioned in Section 2, the functions $k_2(\psi)$ and $d_2(s_1)$ in (3.3), needed to compute the conditional density of s_1 given s_2, can be accurately approximated by the saddlepoint approximation. This problem has recently been reviewed in Pierce and Peters (1992) and will only be described briefly here. Pierce and Peters (1992) is particularly helpful for

clarifying the differences among various formulas in the literature and also provides a careful discussion of the use of the approximation for discrete distributions.

Using the saddlepoint approximation, the conditional log-likelihood function from (3.3) can be approximated by an adjustment to the profile log-likelihood function obtained from the full model (Barndorff-Nielsen and Cox, 1979); that is,

$$l_c(\psi) = \log f(s_1|s_2;\psi) = \psi s_1 - k_2(\psi) - d_2(s_1) \doteq l(\psi, \hat\lambda_\psi) + \tfrac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|, \tag{4.1}$$

where $l(\psi,\lambda) = \psi s_1 + \lambda s_2 - k(\psi,\lambda)$ is the log-likelihood from the joint density of s_1 and s_2, and $j_{\lambda\lambda}(\psi,\lambda) = -\partial^2 l(\psi,\lambda)/\partial\lambda\,\partial\lambda^T$. This approximate conditional likelihood function, which only depends on ψ, can be used in Barndorff-Nielsen's formula (2.1) to provide an approximation to the density of the conditional maximum likelihood estimate:

$$f(\hat\psi_c;\psi) = c|-l_c''(\hat\psi_c)|^{1/2}\exp\{l_c(\hat\psi_c) - l_c(\psi)\}\{1 + O(n^{-3/2})\}.$$

It can also be used in the tail area formula (2.3):

$$F(\hat\psi_c;\psi) = \left\{\Phi(r_c) + \varphi(r_c)\left(\frac{1}{r_c} - \frac{1}{q_c}\right)\right\}\{1 + O(n^{-3/2})\}. \tag{4.2}$$

Here r_c and q_c are the likelihood root and the Wald statistic constructed from $l_c(\psi)$ on the right-hand side of (4.1):

$$r_c = \pm[2\{l_c(\hat\psi_c) - l_c(\psi)\}]^{1/2},$$
$$q_c = (\hat\psi_c - \psi)|-l_c''(\hat\psi_c)|^{1/2}.$$

An equivalent approximation to $F(\hat\psi_c;\psi)$ can be obtained directly from the likelihood function for the joint distribution, and it takes the form

$$F(\hat\psi_c;\psi) = \left\{\Phi(r_p) + \varphi(r_p)\left(\frac{1}{r_p} - \frac{1}{\rho q_p}\right)\right\}\{1 + O(n^{-3/2})\}, \tag{4.3}$$

where r_p is the likelihood root from the profile likelihood function $l_p(\psi) = l(\psi, \hat\lambda_\psi)$,

$$q_p = (\hat\psi - \psi)|-l_p''(\hat\psi)|^{1/2}$$

and

$$\rho = \left\{\frac{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)|}{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|}\right\}^{1/2}.$$

Approximation (4.3) is due to Skovgaard (1987) and is usually called the double saddlepoint approximation (see also Davison, 1988). It also follows immediately from Barndorff-Nielsen (1986), as was shown by DiCiccio and Martin (1991). Approximation (4.2) is derived in Fraser, Reid and Wong (1991) and is called the sequential saddlepoint approximation. The sequential saddlepoint adjusts the likelihood function, as in (4.1), and uses the adjusted version in the one-parameter approximation. The double saddlepoint makes the nuisance parameter adjustment to the tail probability approximation directly, through the factor ρ. In applications the sequential saddlepoint is typically more accurate, although there are exceptions. In the special situation that the function $k_2(\psi)$ in (3.3) is explicitly available, computation of the sequential saddlepoint approximation using the exact conditional likelihood was suggested in Skovgaard (1987).

As at (2.8), writing

$$r^* = r + \frac{1}{r}\log\frac{u}{r},$$

where r and u are either r_c and q_c in (4.2) or r_p and ρq_p in (4.3), we have

$$F(\hat\psi_c;\psi) = \Phi(r^*)\{1 + O(n^{-3/2})\}.$$

4.3 Transformation Models

In this case we seek to approximate the marginal density for one parameter component from the unnormalized joint density for the vector of parameters. This joint density is conditional on the ancillary statistic, as usual, and its unnormalized form is simply proportional to the joint likelihood function.

In evaluating integrals of this type in Bayesian analysis, Tierney and Kadane (1986), Tierney, Kass and Kadane (1989) and Kass, Tierney and Kadane (1988) suggested approximating the integral using Laplace's method. Applying these results shows that the marginal log-likelihood for the parameter of interest can be approximated by adjusting the profile likelihood function:

$$l_m(\psi) = \log f_m(\hat\psi;\psi) = \log\int f(\hat\theta;\theta)\,d\hat\lambda \doteq l(\psi,\hat\lambda_\psi) - \tfrac{1}{2}\log|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|. \tag{4.4}$$

This approximate marginal likelihood can be converted to a tail area approximation using (2.3) and (2.6):

$$F(\hat\psi;\psi) = \left\{\Phi(r_m) + \varphi(r_m)\left(\frac{1}{r_m} - \frac{1}{v_m}\right)\right\}\{1 + O(n^{-3/2})\}, \tag{4.5}$$

where

$$r_m = \pm[2\{l_m(\hat\psi) - l_m(\psi)\}]^{1/2},$$
$$v_m = l_m'(\psi)|-l_m''(\hat\psi)|^{-1/2}.$$

Note the similarity of (4.1) and (4.4), and of (4.2) and (4.5). Result (4.5) was established in DiCiccio, Field and Fraser (1990). It is of the "sequential saddlepoint" form described in Section 4.2, in the sense that the likelihood is adjusted before it is converted to a tail area approximation. There is an equivalent "double saddlepoint" version, also derived in DiCiccio, Field and Fraser (1990), analogous to (4.3); and r^* versions can also be constructed from either of these approximations.

4.4 Discussion

The asymptotic theory outlined in this section and in Section 2.2 has emphasized the use of the signed square root of the likelihood ratio statistic, which requires that the parameter of interest is a scalar. As a referee has stated, since in this case the density approximation given by the p^* formula could be integrated numerically, it is not obvious that tail area approximations (2.3), (2.9), (4.2), (4.3) and (4.5) are particularly useful from a practical point of view. The amount of calculation involved in computing the tail area approximations is still considerably less than numerical integration, though, and where the approximations have been compared to numerical integration the results are generally comparable in accuracy (cf., e.g., Butler, Huzurbazar and Booth, 1992). Use of the tail area approximations is particularly simple for computing p-values for a null hypothesis. As discussed in Sweeting (1995c), the tail area approximations are essentially two-point quadrature formulas, but the two points at which they are evaluated are in some sense specially tailored to the problem. Nevertheless, the surprising accuracy of the tail area approximations is mentioned in many papers in the literature and has not yet been fully explained.

Another reason for emphasizing the tail area approximations, in addition to their apparent accuracy in applications, is that this approach has been quite fruitful for theoretical developments. The similarity between the approximations for exponential families and for transformation families is striking and suggests the possibility of a general approximation formula for tail probabilities. The distinction between the double saddlepoint approximation and the sequential saddlepoint approximation outlined so clearly in Pierce and Peters (1992) highlighted intriguing connections between frequentist and Bayesian inference; see in particular Lindley's (1992) contribution to the discussion, Pierce and Peters' (1992) reply and Pierce and Peters (1994). Further development of these ideas is presented in Sweeting (1995a, b). Another advantage of the tail area approximations is their decreased dependence on the ancillary statistic, as discussed in Section 5.4.

Tail area approximations are, however, only useful for inference about a scalar parameter, and there will be situations when some or all components of a vector parameter are of interest. As a referee has pointed out, it is not clear how the p^* approximation is to be used in this setting. One approach, outlined in Barndorff-Nielsen (1986), is to order the components of a vector parameter in some manner and to construct a series of tail area approximations, one for each component. The same difficulty arises in Bayesian inference and in conditional inference in transformation models, for which a componentwise approach using likelihood ratio statistics is outlined in Fraser and MacKay (1975).

A more usual approach is to base inference for a vector parameter on the log-likelihood ratio statistic, which to first order follows a χ^2_p-distribution, where p is the dimension of the parameter. As has been widely discussed, the statistic is readily adjusted to follow a χ^2_p-distribution to a higher order of approximation, by the technique of Bartlett adjustment. Bartlett adjustment will not be reviewed here; some particularly useful recent references are Barndorff-Nielsen and Cox (1994, Section 4.4, 6.5), DiCiccio and Stern (1993a) and Bickel and Ghosh (1990). The derivation of the Bartlett correction from the p^* formula is given in Barndorff-Nielsen and Cox (1984), where in particular it is shown that the Bartlett-corrected likelihood ratio statistic has the same distribution conditionally and unconditionally. Jensen (1986b) shows that, in the setting of Section 4.2, the Bartlett-corrected likelihood ratio statistic is an $O(n^{-3/2})$ approximation to the conditional distribution of s_1, given the sufficient statistic for the nuisance parameter.

In the scalar parameter setting, a mean and variance adjustment of the signed square root of the log-likelihood ratio statistic is equivalent to a Bartlett adjustment of the log-likelihood ratio statistic itself; in fact Lawley (1956) derived the Bartlett adjustment by considering a sequence of signed likelihood roots. This approach is also used in McCullagh (1987, Section 7.4), Bickel and Ghosh (1990) and Sweeting (1995b).

Although the exponential family versions (4.2) and (4.3) have been fairly thoroughly developed for regression models (Davison, 1988; Pierce and Peters, 1992; Fraser, Reid and Wong, 1991), the results in Section 4.3 have not been systematically applied to regression models of the form $Y = X\beta + \sigma\varepsilon$, all of which can be analyzed using these methods. DiCiccio and Martin (1991, 1993) have extended the range of application of the original formula

somewhat and have illustrated its implementation in several examples. In examples where it has been considered, the double saddlepoint version does not seem to work as well as the sequential saddlepoint version (DiCiccio, Field and Fraser, 1990; DiCiccio and Field, 1991; Butler, Huzurbazar and Booth, 1992).

Conditional models are useful in other contexts for eliminating nuisance parameters or enabling approximate computation of probabilities, or both. These extensions are considered briefly in the next section.

5. EXTENSIONS

5.1 Exponential Families

In many applications, the parameter of interest will not be a linear function of the canonical parameter. In the special case that the parameter of interest is the ratio of two components of the canonical parameter, the nuisance parameter can still be eliminated by conditioning, although some complications arise.

Assume that the density is of the form

$$f(y;\psi,\lambda) = \exp\{\lambda s_1(y) + \psi\lambda s_2(y) - k(\psi,\lambda) - d(y_1,y_2)\},$$

where, as usual, ψ denotes the parameter of interest. For fixed ψ, $s(\psi) = s_1 + \psi s_2$ is sufficient for λ in the ordinary sense. The Jacobian of the transformation from (s_1, s_2) to (s_1, s) is ψ^{-1}, so we have

$$\begin{aligned} f\{s_1|s(\psi)\} &= \psi^{-1}\exp[\lambda s_1 + \lambda(s-s_1) - k(\psi,\lambda) \\ &\qquad - d\{s_1, \psi^{-1}(s-s_1)\}] \\ &\quad\cdot\left[\int \psi^{-1}\exp[\lambda s_1 + \lambda(s-s_1) - k(\psi,\lambda)\right. \\ &\qquad \left. - d\{s_1,\psi^{-1}(s-s_1)\}]\,ds_1\right]^{-1} \\ &= \frac{\exp[-d\{s_1,(s-s_1)/\psi\}]}{\int \exp[-d\{s_1,(s-s_1)/\psi\}]\,ds_1} \\ &= \exp\{d_2(s_1;\psi) - k_2(\psi)\}. \end{aligned} \tag{5.1}$$

Note that the Jacobian of the transformation does not depend on s_1, so it does not affect the conditional distribution. In the terminology of Chamberlin and Sprott (1989), $s(\psi)$ is a linear pivotal. In general, (5.1) is not itself a linear exponential family as was the case in (3.3).

EXAMPLE 5.1. Let s_1 and s_2 be sums of exponential random variables with rate parameters λ and $\psi\lambda$, respectively, based on samples of size n_1 and n_2. Writing $s(\psi)$ for $s_1 + \psi s_2$, it is not hard to show that

$$f(s_1|s(\psi)) = \frac{s_1^{n_1-1}(s-s_1)^{n_2-1}}{\int s_1^{n_1-1}(s-s_1)^{n_2-1}\,ds_1},$$

that is, that $s_1/s(\psi)$ follows a Beta(n_1, n_2) distribution. The marginal distribution of $s(\psi)$ is Gamma(n_1+n_2, λ). If we consider the two exponential random variables to be the waiting times for events in two Poisson processes with rates $\psi\lambda$ and λ, respectively, then $s(\psi)$ corresponds to the total time to n_1+n_2 events from two processes adjusted to have the same rate, λ. From this point of view conditioning on $s(\psi)$ seems quite reasonable.

The result (5.1) suggests a partition of the original joint density as

$$f_1(s_1|s(\psi);\psi) f_2(s(\psi);\psi,\lambda) \tag{5.2}$$

by analogy with (3.2). In Example 5.1 we have the further simplification to

$$f_1(s_1|s(\psi)) f_2(s(\psi);\lambda), \tag{5.3}$$

that is, the ψ-dependence in the joint density is entirely absorbed in the function $s(\psi)$.

A conditional or marginal likelihood cannot be easily defined from the component densities in (5.2), because the differential element associated with the density also depends on ψ. This is true both for $f_1(s_1|s(\psi))$ and for $f_2(s(\psi))$. Thus likelihood-based functions, such as the score function or Fisher information, cannot be obtained in a straightforward manner.

One approach to defining a likelihood function is to consider the problem of testing H_0: $\psi = \psi_0$ and condition on $s(\psi_0) = s_1 + \psi_0 s_2$. This is recommended in Cox and Hinkley (1974, Chapter 5) and McCullagh and Nelder (1989, Chapter 7). The result for the family (5.1) is

$$\begin{aligned} f(s_1|s(\psi_0);\psi,\lambda) &= \exp\left\{\lambda\left(1-\frac{\psi}{\psi_0}\right)s_1 - d\left(s_1,\frac{s-s_1}{\psi_0}\right)\right\} \\ &\quad\cdot\left[\int \exp\left\{\lambda\left(1-\frac{\psi}{\psi_0}\right)s_1\right.\right. \\ &\qquad \left.\left. - d\left(s_1,\frac{s-s_1}{\psi_0}\right)\right\} ds_1\right]^{-1} \\ &= \exp\left\{\lambda\left(1-\frac{\psi}{\psi_0}\right)s_1 - d_s(s_1,\psi_0)\right. \\ &\qquad \left. - c_s(\lambda,\psi,\psi_0)\right\}. \end{aligned} \tag{5.4}$$

Note that this is a linear exponential family for s_1, with canonical parameter $\lambda(1 - \psi/\psi_0)$, but that it is not free of λ unless $\psi = \psi_0$. Because this conditional density has monotone likelihood ratio in s_1 for all λ and all $\psi > \psi_0$ (or $\psi < \psi_0$), uniformly most powerful similar tests can be constructed for testing H_0: $\psi = \psi_0$ and a confidence interval can be built up by testing a succession of ψ_0-values. In Example 5.1 conditioning on $s(\psi_0)$ gives the same conditional density as in (5.3), because the conditional density depends on ψ only through $s(\psi_0)$.

EXAMPLE 5.2. Suppose we have a sample of size n from the gamma distribution and are interested in inference about the mean, which is a ratio of canonical parameters. An expression for $d(s_1, s_2)$ is given in (3.5), although no closed form is available. Thus

$$f(s_1|s(\mu);\mu) = \frac{\exp(-s_1)h(s_1,(s-s_1)\mu)}{\int \exp(-s_1)h(s_1,(s-s_1)\mu)\,ds_1},$$

where $h(s_1, s_2)$ is the volume of the region A defined in (3.5). If we condition on $s(\mu_0)$ instead, we get

$$f(s_1|s(\mu_0);\mu,\lambda) = \exp\left\{\lambda\left(1-\frac{\mu_0}{\mu}\right)s_1 - d_s(s_1,\mu_0) - c_s\left(\frac{\mu_0}{\mu},\lambda\right)\right\}, \tag{5.5}$$

where

$$\exp\{-d_s(s_1,\mu_0)\} = \exp(-s_1)h(s_1,(s-s_1)\mu_0),$$

$$\exp\left\{-c_s\left(\frac{\mu_0}{\mu},\lambda\right)\right\} = \int \exp\{-d_s(s_1,\mu_0)\}\,ds_1.$$

Jensen (1986a) discusses uniform saddlepoint approximations for $h(s_1, s_2)$ to obtain confidence intervals for μ constructed from the conditional density (5.5).

If the parameter of interest in an exponential family model is not either a ratio or a difference of canonical parameters, no exact conditional distribution is available for eliminating the nuisance parameter. While in principle an approximately sufficient statistic could be constructed, no general treatment seems to be available. Constructing an approximate conditional likelihood is a little easier and is discussed in Section 5.3.

5.2 Transformation Models

In transformation models, conditioning on an ancillary gives a dimension reduction to the dimension of the parameter. Nuisance parameters are then eliminated by integrating, that is, by finding the marginal density related to the component of interest.

EXAMPLE 5.3. We generalize the location-regression model (3.7) by writing

$$Y = X\beta + \sigma\varepsilon, \tag{5.6}$$

where $\sigma > 0$ is a scale parameter and ε as before follows a known density $f_0(\cdot)$. The conditional density of $(\hat\beta, \hat\sigma)$, given the maximal ancillary, is given by the general expression (2.1), but in order to eliminate the nuisance parameters by integrating it is necessary to define the pivotals $t_1 = (\hat\beta_1 - \beta_1)/\hat\sigma, \ldots, t_p = (\hat\beta_p - \beta_p)/\hat\sigma$, $t_{p+1} = \hat\sigma/\sigma$; that is,

$$f(\hat\beta,\hat\sigma;\beta,\sigma) \propto f(t_1(\beta_1),\ldots,t_p(\beta_p),t_{p+1}(\sigma)) = f_m(t_1)f_c(t_{(2)}|t_1),$$

where $t_{(2)} = (t_2, \ldots, t_{p+1})$. By analogy with (3.2) the marginal density is free of the nuisance parameters, but as in (5.2) the argument of the density is a pivotal statistic.

A further generalization of interest in the regression model is to allow $f_0(\varepsilon)$ to depend on some shape parameters, say, λ. In this case the full joint density $f(y)$ factorizes as

$$f(y;\beta,\sigma,\lambda) \propto f(t_1(\beta_1),\ldots,t_p(\beta_p),t_{p+1}(\sigma)|d;\lambda)f(d;\lambda), \tag{5.7}$$

where d is the maximal ancillary statistic $(y - X\hat\beta)/\hat\sigma$. This is a generalization of (3.6). Models of this type are discussed in Fraser (1979, Chapter 6) where it is recommended to use $f(d;\lambda)$ for inference about λ and then to use a range of plausible values for λ in constructing inference for (β, σ). A systematic investigation of this approach does not seem to be available in the literature.

5.3 Conditional and Marginal Likelihoods

The results in Sections 5.1 and 5.2 provide expressions for the exact conditional or marginal densities, although not in closed form. As we might hope from earlier sections of the paper, accurate approximations to these densities might be available from saddlepoint approximations or the p^* formula. While very accurate approximations for the density can often be obtained in particular cases, the form of the p^* and r^* formulas suggests that a fruitful approach is to construct approximations from a likelihood function.

As well, in models that are not exponential or transformation families, there is no exact method for eliminating nuisance parameters by conditioning or marginalizing. The only reasonably general non-Bayesian approach to such models seems to be to construct an approximate conditional or marginal

likelihood and to use asymptotic arguments for computing p-values and confidence intervals.

A likelihood construction based on Section 5.1 was suggested in Cox and Reid (1987). For fixed ψ, the restricted maximum likelihood estimate of the nuisance parameter $\hat{\lambda}_\psi$, is at least asymptotically sufficient for λ, so Cox and Reid (1987) suggested computing $f(y|\hat{\lambda}_\psi)$ as the basis for an approximate conditional likelihood. After approximating the marginal density for $\hat{\lambda}_\psi$ by the p^* formula (2.1), the resulting conditional log-likelihood is given by

$$l_{\mathrm{CR}}(\psi) = l(\psi, \hat{\lambda}_\psi) - \tfrac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|, \tag{5.8}$$

where $l(\psi, \lambda)$ is the likelihood function based on the joint density $f(x; \psi, \lambda)$. The construction of $l_{\mathrm{CR}}(\psi)$ follows the conditioning pattern of (5.1); that is, no fixed value ψ_0 of ψ is involved. A version that follows (5.4) was also suggested in Cox and Reid (1987).

In order to compute the conditional density, we need the Jacobian of a one-to-one transformation from y to $(a, \hat{\lambda}_\psi)$, and in (5.8) this has been ignored. An important consequence of this is that $l_{\mathrm{CR}}(\psi)$ is not invariant under one-to-one reparametrizations of the nuisance parameter λ that leave the parameter of interest fixed. Cox and Reid (1987) suggested computing (5.8) using an orthogonal nuisance parameter λ, that is, one satisfying the property $E\{-\partial^2 l(\psi, \lambda)/\partial\psi\,\partial\lambda\} = 0$. Such a parametrization can always be found in the special case that ψ is a scalar.

The lack of invariance can be avoided by using the modified profile likelihood of Barndorff-Nielsen (1983), which is given by

$$l_{\mathrm{BN}}(\psi) = l(\psi, \hat{\lambda}_\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| + \log\left|\frac{d\hat{\lambda}}{d\hat{\lambda}_\psi}\right|, \tag{5.9}$$

where the final term adjusts for the reparametrization in λ. This likelihood was derived in Barndorff-Nielsen (1983) as an approximation to the marginal likelihood of s_2 under the factorization $f(y; \psi, \lambda) = f(s_2; \psi)f(s_1|s_2; \psi, \lambda)$. Several related constructions are discussed in Barndorff-Nielsen (1987). When λ and ψ are orthogonal parameters, the final term in (5.9) is $O(n^{-1})$, whereas the two leading terms are $O(n)$ and $O(1)$, respectively.

All the constructions of (5.9) and (5.8) involve an approximation to the marginal density of $\hat{\lambda}_\psi$, and as noted in Section 5.1 a likelihood constructed from this density is not well defined, because the differential for the density contains ψ-dependence that is ignored. Attempts to address this were made in Kalbfleisch and Sprott (1970) and Fraser and Reid (1989), but no satisfactory resolution seems to be available. Some further points related to this are discussed in Cox (1993).

The derivation of (5.9) in Barndorff-Nielsen (1983) and of (5.8) in Cox and Reid (1987) shows that (5.8) can, at least under parameter orthogonality, provide an approximation to either a conditional or a marginal log-likelihood (ignoring difficulties with the differential mentioned above). In the special cases of the exponential model in Section 4.2 and the transformation model in Section 4.3, where pivotal conditioning or marginalizing is not needed, (5.8) provides an approximation to the exact conditional or marginal log-likelihood to $O(n^{-3/2})$, as described at (4.1) and (4.4). In (4.1) the orthogonal parametrization is defined by $\eta = \partial c(\psi, \lambda)/\partial\lambda$, and the right-hand side of (4.1) becomes $l(\psi, \hat{\eta}_\psi) - (1/2)\log|j_{\eta\eta}(\psi, \hat{\eta}_\psi)|$ after this parameter transformation is made.

Several aspects of (5.8) have been studied recently, and an overview is given in Reid (1992). Although it is often referred to as an approximate conditional likelihood, in many cases it actually approximates the appropriate marginal likelihood as is clear from the construction described in Barndorff-Nielsen (1983). In fact from one point of view it is a simplification of Barndorff-Nielsen's modified profile likelihood. Other simplifications of Barndorff-Nielsen's modified profile likelihood are suggested in McCullagh and Tibshirani (1990), DiCiccio and Stern (1993b) and Barndorff-Nielsen (1994).

A promising use of adjusted likelihoods is in Neyman–Scott–type problems such as described in Example 3.5. Example 3.5 is discussed from this point of view in Barndorff-Nielsen (1983) and in McCullagh and Tibshirani (1990). While the estimate of the common odds ratio from l_{CR} or l_{BN} is not consistent either (as $k \to \infty$), numerical work indicates that it is not as "badly inconsistent" as the usual maximum likelihood estimate. Another example, from Cox and Reid (1992), considers k pairs of independent exponential observations with means $\psi\lambda_j$ and ψ/λ_j, respectively. The maximum likelihood estimate of ψ is $\hat{\psi} = k^{-1}\sum(y_{j1}y_{j2})^{1/2}$, which converges to $\pi\psi/4$. The estimate obtained by maximizing l_{CR} is $(4/3)\hat{\psi}$, which converges to $\pi\psi/3$.

To use $l_{\mathrm{CR}}(\psi)$ or $l_{\mathrm{BN}}(\psi)$ in a tail area formula such as (2.3) or (2.9), it is necessary to understand how it depends on the data, as well as on the parameter, and outside the cases outlined in Sections 4.2 and 4.3 this is fairly difficult. Some progress can be made for the case of a ratio of exponential family parameters, but only by using the linear exponential

structure of (5.4) in an exponential type approximation as in (4.2) (Jensen, 1986b).

However, recent results have shown that the dependence on the data of the log-likelihood (or a conditional or marginal log-likelihood) is needed for approximations accurate to $O(n^{-3/2})$, but not for approximations accurate to $O(n^{-1})$, which is still an improvement on the usual first-order approximations, which are accurate only to $O(n^{-1/2})$. DiCiccio and Martin (1993) describe a tail area approximation derived from l_{CR} accurate to $O(n^{-1})$, and Barndorff-Nielsen and Chamberlin (1994) discuss tail area approximations from l_{BN}, also accurate to $O(n^{-1})$.

5.4 Exponential Regression

Although somewhat specialized, the exponential regression model illustrates several of the points discussed in Sections 4 and 5. Assume that we have a sample of n independent observations $(y_1, \ldots, y_n)$, each following an exponential distribution with mean μ_i. The likelihood function is

$$l(\mu; y) = -\sum \log \mu_i - \mu_i^{-1} \sum y_i.$$

A regression model for the means is assumed to take the generalized linear model form $g(\mu_i) = x_i'\beta$, where x_i is a vector of explanatory variables associated with y_i. We will simplify the following discussion by assuming that there is just one explanatory variable, so that $g(\mu_i) = \beta_0 + \beta_1 x_i$.

If we assume $g(\mu_i) = \mu_i^{-1}$, then

$$l(\beta_0, \beta_1; y) = \sum \log(\beta_0 + \beta_1 x_i) - \beta_0 \sum y_i - \beta_1 \sum y_i x_i$$

is the likelihood function for an exponential family with canonical parameters β_0 and β_1. Inference for β_1, treating β_0 as a nuisance parameter, is constructed from the conditional distribution of $\sum y_i x_i$ given $\sum y_i$. Although exact calculation of this distribution is difficult, the conditional likelihood function can be approximated as in (4.1):

$$l_c(\beta_1) = l(\hat{\beta}_0(\beta_1), \beta_1) + (\tfrac{1}{2}) \log j_{\beta_0\beta_0}(\hat{\beta}_0(\beta_1), \beta_1),$$

where $\hat{\beta}_0(\beta_1)$ is defined by $\sum \{\hat{\beta}_0(\beta_1) + \beta_1 x_i\}^{-1} = \sum y_i$, and $j_{\beta_0\beta_0}(\beta) = \sum (\beta_0 + \beta_1 x_i)^{-2}$. This approximate conditional log-likelihood function can be used in the tail area approximations described in Section 4.2 to construct a confidence interval for β_1 or to test the hypothesis $\beta_1 = 0$. There is no difficulty, in principle, in extending to a multiple regression of the form $\mu_i^{-1} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$; inference for any component, say, β_p, can be constructed from the exact or approximate conditional distribution given $(\sum y_i, \sum y_i x_{1i}, \ldots, \sum y_i x_{p-1,i})$.

In many applications it will be more natural to assume $g(\mu_i) = \log \mu_i$. The log-likelihood function is then

$$l(\beta_0, \beta_1) = -n\beta_0 - \beta_1 \sum x_i - \sum y_i \exp\{-(\beta_0 + \beta_1 x_i)\},$$

which is an exponential family with noncanonical parameters of interest. There is no dimension reduction by sufficiency available, and no obvious ancillary statistic using exponential family theory is available. However, the exponential regression model is a location-regression model on the log scale. Writing $z_i = \log y_i$, we have

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $f(\varepsilon) = \exp(\varepsilon - \exp \varepsilon)$, and the exact density for $(\hat{\beta}_0, \hat{\beta}_1)$ conditional on the ancillary statistic $a = (z_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1, \ldots, z_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)$ is given by the p^* formula, as described in Section 2.1.

The marginal density of $\hat{\beta}_1$ can be obtained exactly as well (cf. Lawless, 1982, Section 6.3) and the log-likelihood corresponding to this exact marginal density is just the profile log-likelihood

$$\begin{aligned} l_m(\beta_1) = l_p(\beta_1) &= -n \log \sum \exp\{a_i + (\hat{\beta}_1 - \beta_1) x_i\} \\ &= -n \log \sum y_i \exp(-\beta_1 x_i). \end{aligned}$$

The same expression for the marginal log-likelihood is obtained from the Laplace approximation (4.4), which in this case is exact.

The approximate conditional likelihood of Cox and Reid (1987) is readily computed. Centering the x's (i.e., assuming $\sum x_i = 0$) ensures that the parameters β_0 and β_1 are orthogonal, and in fact $j_{\beta_0\beta_0}(\hat{\beta}_0(\beta_1), \beta_1) = n$, which does not depend on β_1. As a result $l_{CR}(\beta_1)$ is the same as the profile log-likelihood. It is easily shown as well that $l_{BN}(\beta_1)$ is equal to the profile likelihood (Barndorff-Nielsen and Cox, 1994, Example 8.8). However, if $l_{CR}(\beta_1)$ is computed using the nuisance parametrization $\lambda = \exp \beta_0$, then it is equal to $\{(n-1)/n\} l_p(\beta_1)$ (Barndorff-Nielsen and Cox, 1994, Example 4.8), highlighting the potential difficulty with the lack of invariance of $l_{CR}(\cdot)$. No such difficulty arises with $l_{BN}(\cdot)$.

Extending the regression for z_i to a multiple regression is again straightforward, and the exact marginal log-likelihood for $(\beta_1, \ldots, \beta_p)$ having integrated out β_0 is again given by the profile log-likelihood. Further exact or approximate integration is required for inference about a component parameter; alternatively multiparameter methods mentioned in Section 4.4 can be used for inference about some or all the regression coefficients.

Introducing an additional shape parameter into the model for y_i can be done in a variety of ways; the

two most obvious being Weibull and gamma models. For example, if we assume $y_i \sim \Gamma(\nu, \mu_i)$ and we use the canonical link $\mu_i^{-1} = \beta_0 + \beta_1 x_i$, the resulting exponential family log-likelihood is

$$l(\beta_0, \beta_1, \nu; y) = -\nu\beta_0 s_1 - \nu\beta_1 s_2 + \nu s_3 + \nu \sum \log(\beta_0 + \beta_1 x_i) + n\nu \log \nu - n \log \Gamma(\nu),$$

where $s_1 = \sum y_i$, $s_2 = \sum x_i y_i$ and $s_3 = \sum \log y_i$. The shape parameter ν is a component of the canonical parameter, and the method outlined in Example 3.3 and Section 4.2 applies directly. The regression parameter β_1 is a ratio of canonical parameters, and exact conditonal inference can be constructed as described in Section 5.2. Computation of the approximate conditional likelihood is straightforward.

In the model with the link function $\log \mu_i = \beta_0 + \beta_1 x_i$, a shape parameter can be introduced by assuming a Weibull distribution for y_i, or equivalently by assuming a model of the form (5.6) for $z_i = \log y_i$, with $f(\varepsilon)$ the extreme value distribution given above, and the discussion of Example 5.3 applies directly. Inference based on the generalized gamma distribution, for which the Weibull is a special case, is discussed in Lawless (1982, Section 6.6), and provides an example of the extension described by (5.7).

5.5 Approximate Sufficiency and Ancillarity

It would seem quite natural in the light of the development of higher order approximations to consider local ancillarity and sufficiency in the presence of nuisance parameters. For example, in the pattern of factorization (3.2), if $f(s_1|s_2; \psi, \lambda_0 + \varepsilon n^{-1/2}) = f(s_1|s_2; \psi, \lambda_0)\{1 + O(n^{-1})\}$, we could say that s_2 is second-order locally sufficient for λ. Similarly, if $f(s_2; \psi_0 + \delta n^{-1/2}, \lambda) = f(s_2; \psi_0, \lambda)\{1 + O(n^{-1})\}$, s_2 is second-order locally ancillary for ψ. An approach to defining local cuts this way is described in Christensen and Kiefer (1994). These ideas are also considered in Severini (1993, 1994a), using slightly different definitions from those given here. In Severini (1994a) the dependence of the distribution of the conditioning statistic s_2 is defined via local power of a certain test statistic, rather than by requiring its distribution to be approximately free of the parameter of interest.

In Barndorff-Nielsen (1986), the result

$$f(\hat{\lambda}_\psi, r_\psi^*; \psi, \lambda) = \varphi(r_\psi^*) p^*(\hat{\lambda}_\psi | r_\psi^*; \lambda)\{1 + O(n^{-3/2})\} \tag{5.10}$$

is established, where $\hat{\lambda}_\psi$ is the restricted maximum likelihood estimate of the nuisance parameter, and r_ψ^* is the modified likelihood ratio statistic computed from the profile likelihood, via formula (2.8). [See, in particular, his equations (3.6) and (3.12). A simpler expression than (3.12) for u is given in Barndorff-Nielsen (1991).] This result is similar to the complementary expression to that given at (5.3); that is, the joint density factors as $f(s_1|s(\psi); \lambda) f(s(\psi))$ and is an extension of the result of McCullagh (1984) on the local sufficiency of the likelihood root.

Barndorff-Nielsen's result (5.10) assumes that in the full model $f(y; \psi, \lambda)$ it is possible to find an approximate ancillary statistic so that the p^* formula for the distribution of $(\hat{\psi}, \hat{\lambda})$ holds to order $O(n^{-3/2})$, as described briefly in Section 2.3. A general construction of such ancillaries, suitable for approximating tail probabilities, is given in Fraser and Reid (1995). The same construction is used in the submodel obtained by fixing ψ to find an approximate ancillary for the nuisance parameter λ, and the marginal distribution of this statistic is used for inference about ψ. It is easier to find an approximate ancillary for use in tail area approximations, as it is only necessary to find the direction in which to take the sample space derivative of the log-likelihood function [cf. (2.5)], and this does not require complete specification of the ancillary statistic. This is one advantage that tail area approximations have over the p^* approximation.

As a referee has pointed out, (5.10) gives a third-order result based on the profile likelihood, rather than any adjusted version of profile likelihood (cf. the double saddlepoint approximation of Section 4.2). Approximations to $O(n^{-1})$ can be obtained from the profile likelihood that do not require specification of an approximate ancillary; several such are discussed in DiCiccio and Martin (1993) and Barndorff-Nielsen and Chamberlin (1994). However, numerical work in those papers and in Pierce and Peters (1992) does suggest that tail area approximations based on adjusted log-likelihoods are more accurate.

5.6 Information and Estimating Equations

An approach to justifying conditional inference under factorization (3.2) and marginal inference (for λ) under (3.6) is to try to show that the information about the parameter of interest contained in the conditional or marginal density that is ignored is negligible. Several methods of quantifying this information were presented in Barndorff-Nielsen (1978, Chapter 4). These are reviewed and extended in Jorgensen (1993).

When the conditional or marginal statistic of interest is a pivotal, and consequently the conditional or marginal likelihood is not well defined, it is also

not clear how to define quantities derived from the likelihood, such as Fisher information. The problem of defining information in a pivotal quantity also arises in the theory of estimating equations, which are by definition functions of the data and the parameter. The information in an estimating equation is usually defined by projection (Godambe, 1980; Liang, 1983; Efron, 1977; Bhapkar, 1991).

The construction of estimating equations in the presence of nuisance parameters typically assumes a factorization of the form (3.2) or (5.2). Godambe (1976) established the optimality of the score equation from $f(s_1|s_2;\psi)$ in the setting expressed in (3.2), where there exists a sufficient statistic for the nuisance parameter. His optimality result did not entail conditions on the amount of information in the marginal density $f(s_2;\psi,\lambda)$, but did entail conditions on the class of estimating equations. It is an estimating equation analogue of the result mentioned in Section 3 that the class of tests with size not depending on λ must be constructed from the conditional distribution $f(s_1|s_2;\psi)$. Both results require the family of densities for s_2 to be complete.

Lindsay (1982, 1983) considered optimality in the more general setting where the sufficient statistic for λ depends on the parameter of interest. The optimal score function is obtained from $f(s_1|s(\psi_0))$ as in (5.4) and typically depends on λ. Recent work in Lindsay and Waterman (1992) considers approximations to the optimal score function. A review of the estimating function approach to eliminating nuisance parameters is given in Liang and Zeger (1995).

6. OTHER ROLES OF CONDITIONING

6.1 Conditioning for Convenience

Conditional distributions are much easier to compute than marginal distributions, and some recent developments have used conditional distributions either in place of or as approximations to marginal distributions.

Possibly the most fruitful example of this is the use of sequences of conditional densities in Markov chain Monte Carlo methods to generate samples from desired joint and/or marginal distributions. A thorough survey is given in Besag, Green, Higdon and Mengersen (1995). This approach has been especially useful in Bayesian inference for computing marginal posteriors in high-dimensional problems. An introductory account is given in Casella and George (1992).

An earlier example of conditioning for convenience is the construction of the partial likelihood for the proportional hazards model (Cox, 1972, 1975). This is obtained from a product of conditional distributions, each of which is free of the nuisance parameter. A similar approach was suggested in Besag (1975) to construct a pseudolikelihood for spatial models. A survey of examples and efficiency results for these and similar likelihoods is given in Lindsay (1988).

Fraser and Massam (1985) suggested testing hypotheses about a vector parameter, by means of conditioning for convenience, as follows. To test $H\colon\ \theta = \theta_0$ in the model $f(y;\theta)$, let $U(\theta_0)$ be the score vector $\partial \log f(y;\theta_0)/\partial\theta$ and $||U(\theta_0)||$ its (Euclidean) length. An observed level of significance is defined as

$$p(\theta_0) = \Pr\left\{||U(\theta_0)|| \geq u_0 \middle| \frac{U}{||U(\theta_0)||}\right\} \tag{6.1}$$

and measures the probability of a departure as large as or larger than the observed departure in magnitude, conditional on the direction of departure. Skovgaard (1988) provides a Lugannani–Rice–type approximation which can be applied to this problem. Note that the departure of the data from the hypothesized value is measured in only one direction: that indicated by the observed data value. Fraser and Massam (1985) referred to these tests as conical tests; the name directional was suggested in Skovgaard (1988). Cheah, Fraser and Reid (1994) apply this idea to multiparameter testing in exponential models.

If a conditional density turns out not to depend on the conditioning event, then it is of course a marginal density. In Fraser, Lee and Reid (1990) this idea is exploited to obtain an appproximation to the marginal density of the t-pivotal in regression model (5.6). A conditional density is constructed which is approximately constant as a function of the conditioning variable. Fraser, Lee and Reid (1990) discuss how the dependence of the approximating conditional density can be assessed by a Monte Carlo procedure.

Kolassa and Tanner (1994) combine asymptotic theory and the Gibbs sampler by using Skovgaard's (1987) saddlepoint approximation to the conditional distribution of interest and then constructing a Markov chain by sampling from this approximation to the conditional distribution. Skovgaard's approximation is a double saddlepoint approximation, as described in Section 4.2. The sequential saddlepoint approximation is usually more accurate, but perhaps too computationally cumbersome to use as part of the Markov chain Monte Carlo algorithm.

Although conditional distributions are easier to compute, not requiring evaluation of high-dimensional integrals, they are not easy to simulate

from, and simulation-based inference is arguably easier to do unconditionally. An approach to conditional simulation tailored to a specific example is outlined in Davison and Hinkley (1988). A systematic approach is described in detail in Booth, Hall and Wood (1992).

6.2 Conditioning and Power

A point often made in discussions of conditional inference is that conditional tests are less powerful than unconditional tests. An interesting simple example is given in Cox and Hinkley (1974, Example 4.6). This point is often raised in discussions of the 2×2 table, although there the problem is largely caused by the discreteness of the sample space. In a discrete sample space, the set of achievable α-levels or p-values will be smaller for a conditional test than for an unconditional test, because the conditioning defines a partition of the sample space, and there are fewer points in the subspace defined by the conditioning than there are in the full sample space. This is not a problem with conditioning per se, but rather with the interpretation given to p-values. As discussed in Agresti (1992), it seems likely that the problem can be largely alleviated by the use of some smoothing of p-values.

Even in continuous sample spaces, though, a given conditional procedure may be less powerful than an unconditional procedure. While this is relevant to a long-run assessment of the procedure over the variety of circumstances modelled by the full sample space, it does not seem particularly relevant in individual instances. Of course the usual Neyman–Pearson–type arguments can be used in principle to construct conditional tests that are exactly or approximately optimal among the class of all conditional tests.

A related issue is the conditional properties of a given unconditional procedure. From the point of view of the approximation theory considered in this paper, this question is addressed in Severini (1990). More generally, there is a large literature on investigating the conditional properties of a given procedure by establishing the existence or nonexistence of relevant subsets of the sample space, that is, subsets determined typically by an ancillary statistic, within which an unconditional property such as a confidence limit is invalid. This approach is reviewed in Casella and Goutis (1995).

6.3 Conditional and Bayesian Inference

Conditioning on ancillary statistics or on sufficient statistics for nuisance parameters, in addition to being convenient for the implementation of accurate approximations, is related to the idea of making long-run frequency properties of inference statements such as significance levels or confidence limits relevant to an individual instance. The Bayesian approach avoids the dilemma of maintaining a long-run frequency interpretation while looking for relevance in a particular instance, although one could argue that the problem has simply been shifted from establishing relevance to that of choosing a prior. This latter problem is similarly difficult to solve purely by mathematical techniques.

On a more technical level, in transformation models Bayesian inference with "flat" priors is the same as inference from the conditional distribution given the ancillary statistic (see, e.g., Lawless, 1982, Appendix G); that is, the Bayes posterior interval with probability $1 - \alpha$, say, is exactly the same as the $1 - \alpha$ confidence interval obtained from the conditional density. This is implicit in expression (2.2) and its generalization to transformation models.

As expression (2.2) is a restatement of the p^* formula for location models, one might expect that approximate confidence intervals obtained from p^* and r^* formulas in general models have an approximate Bayesian interpretation as well, and this has been investigated in a number of recent papers. From one point of view the goal is to identify priors for which the Bayesian posterior distribution is also an accurate frequentist solution (Welch and Peers, 1963; Stein, 1985; Tibshirani, 1989; Nicolau, 1993). Another aspect is that p^* does itself have a Bayesian interpretation (Davison, 1986); and, as was pointed out in Barndorff-Nielsen (1987) and Sweeting (1987), the approximate conditional likelihood (5.8) can also be derived as an approximation to the log of the marginal posterior density for ψ by using a Laplace approximation and assuming the priors for ψ and λ are independent or that the prior for λ given ψ is free of ψ, at least near $\hat{\lambda}_\psi$. A systematic approach to the relationship between Bayesian and frequentist asymptotics is outlined in Sweeting (1995a, b). An interesting approach to approximate ancillarity based on Bayesian inference is proposed in Severini (1994b): a statistic is called Bayes-ancillary if the posterior distribution for θ is the same in the conditional and unconditional models.

The foundational aspects of conditioning have been much debated. Particular emphasis is often given to Birnbaum's theorem (Birnbaum, 1962) that the principles of conditioning on ancillary statistics and marginalizing to sufficient statistics imply that all the information that the data can offer is summarized in the likelihood function. In fact the so-called conditionality principle alone entails the

likelihood principle (Evans, Fraser and Monette, 1986). Berger and Wolpert (1984) discuss these results in some detail and argue that the most effective strategy for implementing the likelihood principle is a Bayesian approach. A relatively recent discussion of some foundational issues is given in Brown (1990) and the discussion therein.

7. CONCLUSION

It is not possible to cover in any depth all the roles of conditioning in inference, and, in particular, Section 6 is quite sketchy and incomplete. The close relationship between conditional distributions and developments in the theory of higher-order asymptotics have been very fruitful for the theory of inference, and I have tried to outline what I think are the essential ideas in this relationship.

There are certainly many new things still to be learned, particularly in connection with approximate ancillarity and sufficiency in general models, and properties of approximate conditional and marginal likelihoods. At the same time, the existence of relatively accurate approximations for fairly wide classes of problems will have an impact on statistical methodology, particularly as computer programs become available for easy implementation.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge helpful discussions with D. R. Cox, D. A. S. Fraser and T. J. DiCiccio. Extremely useful comments on an earlier version were provided by T. A. Severini, R. E. Kass and a referee. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

AGRESTI, A. (1992). A survey of exact inference for contingency tables (with discussion). *Statist. Sci.* **7** 131–177.

ANDERSEN, E. B. (1973). *Conditional Inference and Models for Measuring*. Mentalhygiejnisk Forlag, Copenhagen.

BARNARD, G. A. (1984). Discussion of "Tests of significance for 2×2 contingency tables" by F. Yates. *J. Roy. Statist. Soc. Ser. A* **147** 449–450.

BARNDORFF-NIELSEN, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

BARNDORFF-NIELSEN, O. E. (1980). Conditionality resolutions. *Biometrika* **67** 293–310.

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

BARNDORFF-NIELSEN, O. E. (1984). On conditionality resolution and the likelihood ratio for curved exponential models. *Scand. J. Statist.* **11** 157–170. [Correction: *Scand. J. Statist.* **12** (1985) 190.]

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. (1987). Discussion of "Parameter orthogonality and approximate conditional inference" by D. R. Cox and N. Reid. *J. Roy. Statist. Soc. Ser. B* **49** 18–19.

BARNDORFF-NIELSEN, O. E. (1988a). Discussion of "Saddlepoint methods and statistical inference" by N. Reid. *Statist. Sci.* **3** 228–229.

BARNDORFF-NIELSEN, O. E. (1988b). *Parametric Statistical Models and Likelihood*. Springer, New York.

BARNDORFF-NIELSEN, O. E. (1990). Approximate interval probabilities. *J. Roy. Statist. Soc. Ser. B* **52** 485–496.

BARNDORFF-NIELSEN, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78** 557–564.

BARNDORFF-NIELSEN, O. E. (1994). Adjusted versions of profile likelihood and likelihood roots, and extended likelihood. *J. Roy. Statist. Soc. Ser. B* **56** 125–140.

BARNDORFF-NIELSEN, O. E. and CHAMBERLIN, S. R. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81** 485–499.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 279–312.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. Roy. Statist. Soc. Ser. B* **46** 483–495.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72** 355–366.

BASU, D. (1978). On partial sufficiency: a review. *J. Statist. Plann. Inference* **2** 1–13.

BERGER, J. O. and WOLPERT, R. L. (1984). *The Likelihood Principle*. IMS, Hayward, CA.

BESAG, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10** 3–66.

BHAPKAR, V. P. (1991). Sufficiency, ancillarity and information in estimating functions. In *Estimating Functions* (V. P. Godambe, ed.) 241–254. Oxford Univ. Press.

BICKEL, P. J. and GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and Bartlett correction—a Bayesian argument. *Ann. Statist.* **18** 1070–1090.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57** 269–306.

BOOTH, J., HALL, P. and WOOD, A. (1992). Bootstrap estimation of conditional distributions. *Ann. Statist.* **20** 1594–1610.

BRESLOW, N. E. (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68** 73–84.

BRESLOW, N. E. and DAY, N. (1980). *Statistical Methods in Cancer Research, 1: The Analysis of Case-control Studies*. IARC, Lyon.

BROWN, L. D. (1990). An ancillarity paradox in multiple regression. *Ann. Statist.* **18** 471–533.

BUTLER, R. W., HUZURBAZAR, S. and BOOTH, J. G. (1992). Saddlepoint approximations for the Bartlett–Nanda–Pillai trace statistic in multivariate analysis. *Biometrika* **79** 705–716.

CASELLA, G. and GEORGE, E. I. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46** 167–174.

CASELLA, G. and GOUTIS, C. (1995). Frequentist post-data inference. *Internat. Statist. Rev.* To appear.

CHAMBERLIN, S. A. and SPROTT, D. A. (1989). Linear systems of pivotals and associated pivotal likelihoods with applications. *Biometrika* **76** 685–691.

CHEAH, P. K., FRASER, D. A. S. and REID, N. (1994). Multiparameter testing in exponential models: third order approximations from likelihood. *Biometrika* **81** 271–278.

CHRISTENSEN, B. J. and KIEFER, N. M. (1994). Local cuts and separate inference. *Scand. J. Statist.* **21** 389–402.

COX, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

COX, D. R. (1958b). The regression analysis of binary sequences (with discussion). *J. Roy. Statist. Soc. Ser. B* **20** 215–242.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

COX, D. R. (1980). Local ancillarity. *Biometrika* **67** 279–286.

COX, D. R. (1988). Some aspects of conditional and asymptotic inference. *Sankhyā Ser. A* **50** 314–337.

COX, D. R. (1993). Unbiased estimating equations derived from statistics that are functions of a parameter. *Biometrika* **80** 905–909.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

COX, D. R. and REID, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79** 408–411.

DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631–650.

DANIELS, H. E. (1958). Discussion of "The regression analysis of binary sequences" by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **20** 236–238.

DANIELS, H. E. (1987). Tail probability approximations. *Internat. Statist. Rev.* **55** 37–48.

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73** 323–332.

DAVISON, A. C. (1988). Approximate conditional inference in generalized linear models. *J. Roy. Statist. Soc. Ser. B* **50** 445–461.

DAVISON, A. C. and HINKLEY, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75** 417–431.

DAWID, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 79–110.

DICICCIO, T. J. and FIELD, C. A. (1991). An accurate method for approximate conditional and Bayesian inference about linear regression models from censored data. *Biometrika* **78** 903–910.

DICICCIO, T. J., FIELD, C. A. and FRASER, D. A. S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77** 77–95.

DICICCIO, T. J. 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.

DICICCIO, T. J. and MARTIN, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. Ser. B* **55** 305–316.

DICICCIO, T. J. and STERN, S. E. (1993a). On Bartlett adjustments for approximate Bayesian inference. *Biometrika* **80** 731–740.

DICICCIO, T. J. and STERN, S. E. (1993b). An adjustment to profile likelihood based on observed information. Technical Report 424, Dept. Statistics, Stanford Univ.

DURBIN, J. (1980a). Approximations for densities of sufficient estimators. *Biometrika* **67** 311–333.

DURBIN, J. (1980b). The approximate distribution of serial correlation coefficients. *Biometrika* **67** 334–349.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65** 457–487.

EVANS, M. J., FRASER, D. A. S. and MONETTE, G. (1986). On principles and arguments to likelihood (with discussion). *Canad. J. Statist.* **14** 181–200.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. London Ser. A* **134** 285–307.

FRASER, D. A. S. (1956). Sufficient statistics with nuisance parameters. *Ann. Math. Statist.* **27** 838–842.

FRASER, D. A. S. (1968). *The Structure of Inference*. Wiley, New York.

FRASER, D. A. S. (1979). *Inference and Linear Models*. McGraw-Hill, New York.

FRASER, D. A. S. (1988). Normed likelihood and saddlepoint approximation. *J. Multivariate Anal.* **27** 181–193.

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

FRASER, D. A. S., LEE, H. S. and REID, N. (1990). Nonnormal linear regression: an example of significance levels in high dimensions. *Biometrika* **77** 333–341.

FRASER, D. A. S. and MACKAY, J. (1975). Parameter factorization and inference based on significance likelihood and objective posterior. *Ann. Statist.* **3** 559–572.

FRASER, D. A. S. and MASSAM, H. (1985). Conical tests: observed levels of significance and confidence regions. *Statist. Hefte* **26** 1–18.

FRASER, D. A. S. and REID, N. (1989). Adjustments to profile likelihood. *Biometrika* **76** 477–488.

FRASER, D. A. S. and REID, N. (1995). Ancillaries and third order significance. *Utilitas Math.* **47** 33–54.

FRASER, D. A. S., REID, N. and WONG, A. (1991). Exponential linear models: a two-pass procedure for saddlepoint approximation. *J. Roy. Statist. Soc. Ser. B* **53** 483–492.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63** 277–284.

GODAMBE, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika* **67** 155–162.

HINKLEY, D. V. (1980a). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

HINKLEY, D. V. (1980b). Likelihood. *Canad. J. Statist.* **8** 151–164.

JENNISON, C. (1992). Discussion of "Practical use of higher order asymptotics for multiparameter exponential families" by D. A. Pierce and D. Peters. *J. Roy. Statist. Soc. Ser. B* **54** 728.

JENSEN, J. L. (1986a). Inference for the mean of a gamma distribution with unknown shape parameter. *Scand. J. Statist.* **13** 135–151.

JENSEN, J. L. (1986b). Similar tests and the standardized log likelihood ratio statistic. *Biometrika* **73** 567–572.

JENSEN, J. L. (1992). The modified signed log likelihood statistic and saddlepoint approximation. *Biometrika* **79** 693–704.

JORGENSEN, B. (1993). The rules of conditional inference: is there a universal definition of nonformation? In *Proceedings of the 49th Session of the ISI* 323–340. Internat. Statist. Inst., Voorburg.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. Ser. B* **32** 175–208.

KASS, R. E. (1989). The geometry of asymptotic inference (with discussion). *Statist. Sci.* **4** 188–234.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1988). Asymptotics in Bayesian computation. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 261–278. Oxford Univ. Press.

KOLASSA, J. E. and TANNER, M. A. (1994). Approximate conditional inference in exponential families via the Gibbs sampler. *J. Amer. Statist. Assoc.* **89** 697–702.

LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

LAWLEY, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika* **43** 295–303.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

LIANG, K.-Y. (1983). On information and ancillarity in the presence of a nuisance parameter. *Biometrika* **70** 607–612.

LIANG, K.-Y. (1987). Estimating functions and approximate conditional likelihood. *Biometrika* **74** 695–702.

LIANG, K.-Y. and ZEGER, S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statist. Sci.* **10** 158–173.

LINDLEY, D. V. (1992). Discussion of "Practical use of higher order asymptotics for multiparameter exponential families" by D. A. Pierce and D. Peters. *J. Roy. Statist. Soc. Ser. B* **54** 728.

LINDSAY, B. G. (1982). Conditional score functions: some optimality results. *Biometrika* **69** 503–512.

LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

LINDSAY, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80** 221–239.

LINDSAY, B. G. and WATERMAN, R. P. (1992). Extending Godambe's method in nuisance parameter problems. In *Proceedings of a Symposium in Honour of V. P. Godambe*. Dept. Statistics, Univ. Waterloo.

LUGANNANI, R. and RICE, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.* **12** 475–490.

MCCULLAGH, P. (1984). Local sufficiency. *Biometrika* **71** 233–244.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MCCULLAGH, P. and TIBSHIRANI, R. J. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **52** 325–344.

NICOLAU, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. Ser. B* **55** 377–390.

PIERCE, D. A. and PETERS, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 701–738.

PIERCE, D. A. and PETERS, D. (1994). Higher order asymptotics and the likelihood principle: one parameter models. *Biometrika* **81** 1–10.

PLACKETT, R. (1977). The marginal totals of a 2 × 2 table. *Biometrika* **64** 37–42.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366.

REID, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statist. Sci.* **3** 213–238.

REID, N. (1992). Aspects of modified profile likelihood. In *Proceedings of the International Symposium on Nonparametric Statistics and Related Topics* (A. K. Saleh, ed.) 423–442. Elsevier, Amsterdam.

SEVERINI, T. (1990). Conditional properties of likelihood based significance tests. *Biometrika* **77** 343–352.

SEVERINI, T. (1993). Local ancillarity in the presence of a nuisance parameter. *Biometrika* **80** 305–320.

SEVERINI, T. (1994a). On the approximate elimination of nuisance parameters by conditioning. *Biometrika* **81** 649–661.

SEVERINI, T. (1994b). Information and conditional inference. Technical report, Dept. Statistics, Northwestern Univ.

SKOVGAARD, I. M. (1985). A second order investigation of asymptotic ancillarity. *Ann. Statist.* **13** 534–551.

SKOVGAARD, I. M. (1986). Successive improvement of the order of ancillarity. *Biometrika* **73** 516–519.

SKOVGAARD, I. M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.* **24** 875–887.

SKOVGAARD, I. M. (1988). Saddlepoint expansions for directional test probabilities. *J. Roy. Statist. Soc. Ser. B* **50** 269–280.

SKOVGAARD, I. M. (1990). On the density of minimum contrast estimators. *Ann. Statist.* **18** 779–789.

STEIN, C. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*. PWN–Polish Scientific Publishers, Warsaw.

SWEETING, T. J. (1987). Discussion of "Parameter orthogonality and approximate conditional inference" by D. R. Cox and N. Reid. *J. Roy. Statist. Soc. Ser. B* **49** 20–21.

SWEETING, T. J. (1995a). A Bayesian approach to approximate conditional inference. *Biometrika* **82** 25–36.

SWEETING, T. J. (1995b). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82** 1–24.

SWEETING, T. J. (1995c). Approximate Bayesian computation based on signed roots of log-density ratios. In *Bayesian Statistics 5* (J. O. Berger, J. M. Bernardo, D. V. Lindley and A. F. M. Smith, eds.). Oxford Univ. Press. To appear

TIBSHIRANI, R. J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76** 604–608.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–87.

TIERNEY, L. J., KASS, R. E. and KADANE, J. B. (1989). Approximation of marginal densities of nonlinear functions. *Biometrika* **76** 425–433. [Correction: *Biometrika* **78** (1991) 233–234.]

WELCH, B. L. and PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **25** 318–329.

YATES, F. (1984). Tests of significance for 2 × 2 tables (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 426–463.

ZHU, Y.-L. and REID, N. (1994). Information, ancillarity and sufficiency in the presence of nuisance parameters. *Canad. J. Statist.* **22** 111–124.