International Statistical Review

International Statistical Review (2017), 85, 3, 519–541 doi:10.1111/insr.12232

Modern Likelihood-Frequentist Inference

Donald Alan Pierce[1] and Ruggero Bellio[2]

[1]*Statistics Department, Oregon State University, Corvallis, OR, USA*

[2]*Dipartimento di Scienze Economiche e Statistiche, Università di Udine, Udine, Italy*
E-mail: ruggero.bellio@uniud.it

Summary

We offer an exposition of modern higher order likelihood inference and introduce software to implement this in a quite general setting. The aim is to make more accessible an important development in statistical theory and practice. The software, implemented in an `R` package, requires only that the user provide code to compute the likelihood function and to specify extra-likelihood aspects of the model, such as stopping rule or censoring model, through a function generating a dataset under the model. The exposition charts a narrow course through the developments, intending thereby to make these more widely accessible. It includes the likelihood ratio approximation to the distribution of the maximum likelihood estimator, that is the p^* formula, and the transformation of this yielding a second-order approximation to the distribution of the signed likelihood ratio test statistic, based on a modified signed likelihood ratio statistic r^*. This follows developments of Barndorff-Nielsen and others. The software utilises the approximation to required Jacobians as developed by Skovgaard, which is included in the exposition. Several examples of using the software are provided.

Key words: Ancillary statistic; conditional inference; likelihood asymptotics; modified profile likelihood; modified signed likelihood ratio; neo-Fisherian inference; p^* formula; saddlepoint approximation.

1 Introduction and Basic Concepts

1.1 Introductory Aims of Paper

Special likelihood-based procedures, modifying usual inferential approximations for much higher accuracy, have emerged in recent years; see glimpses provided by Davison (2003, Chapter 12), Brazzale & Davison (2008), Lozada-Can & Davison (2010). The performance of these methods is superb and often close to exact. These improvements are of practical interest, for example as seen in the examples here, where P-values near 0.05 are often changed by a factor of 2. However, there is considerably more to modern likelihood asymptotics than this numerical accuracy, in terms of extended (or neo-) Fisherian inference, as treated to a limited extent here. This paper includes both an exposition aiming to make more accessible the main ideas of this development and provides novel software tools for implementing them.

Modern likelihood inference has much to do with higher order asymptotics, particularly second order $O(n^{-1})$ as contrasted with the usual first order $O(n^{-1/2})$. We emphasise that the higher order versions are often suitable for small samples, the convergence in sample size being far more rapid.

After dealing with preliminary issues in Section 2, Sections 3–4 present the main higher order theory, for approximating the distribution of parameter estimates and transforming from this to the distribution of the likelihood ratio test statistic. Section 5 deals with approximation of Jacobians required for the latter step. The challenge of this step hindered the development of the main theory for the decade 1986–1996. This was because the Jacobians are likelihood partial derivatives with respect to parameter estimates, holding fixed ancillary statistics that can be largely notional. Section 6 pertains to discrete data and to similar tests in exponential families.

The tools for applying this are embodied in the package `likelihoodAsy` in `R` (R Core Team, 2017), which is available at the Comprehensive R Archive Network. This package applies quite generally—well beyond independent observations, exponential families and transformation models, requiring primarily only a user-supplied function for evaluating the likelihood. Models where evaluating the likelihood requires numerical integration are exemplified. Inferences beyond first order require model specification beyond the likelihood function, such as stopping rules or censoring models, which is achieved by another user-provided function that generates a sample under the model.

This raises interesting issues about asymptotic methods. Suppose data from sequential Bernoulli trials with success probability p yields r successes in n trials. Exact frequentist inference depends on whether one has stopped on the r-th failure or on the n-th trial. Under mild conditions, the likelihood function is in either case proportional to $p^y\,(1-p)^{n-y}$. With many first-order asymptotic methods, the inference depends only on the likelihood function, while higher order methods will better approximate the exact inference and thus depend on more than the likelihood function. Cox and Hinkley (1974, example 2.34) give one characterisation of stopping rules that do not affect the likelihood function. Similar issues also arise in terms of specifying the mechanism giving rise to censored data. Kalbfleisch and Prentice (2002, section 5.2) discuss censoring models that do not affect the likelihood function. Of course, the gain in this respect of higher order methods can depend on such models being correctly formulated, which is not a lack of robustness. It may be uncommon to desire inference in terms of models for censoring—these are seldom intended to be realistic; see Pierce & Bellio (2015). On the other hand, there is considerable applied interest in the effect of stopping rules, for example in sequential clinical trials, and higher order methods are useful in studying this; see Pierce & Bellio (2006). We return to these matters following Eq. (13) in Section 4.

We have particular motivations for providing the exposition for non-experts of the basis for this theory, which is not intended as a complete review or survey of its development, as was provided by Reid (1988, 1996, 2003). It is fair to predict that the development over the past 30 years is nearly complete. The original developments were in more advanced and esoteric terms than meet the needs for a widespread grasp of them, as might find its way into textbooks; for example see Young & Smith (2005). Without this wider dissemination, our concern is that this important chapter in the theory of inference will largely fade away following the final stages of development. Making more accessible the main ideas requires carefully choosing a narrow path through the developments. As is typical for major advances, it is possible in retrospect to describe the main ideas and results much more simply, which is our aim. Others have chosen, usefully, to make such exposition quite differently from this, for example Brazzale *et al.* (2007), Brazzale & Davison (2008) and Lozada-Can & Davison (2010). Accompanying the text by Brazzale, Davison and Reid were software tools of a different nature than here.

There are largely two limiting issues regarding the adequacy of first-order methods: (i) limited information on the interest parameter and (ii) effects of fitting nuisance parameters. Issue (i) is the 'small sample size' matter that would first come to mind in considering adequacy of asymptotics but (ii) can be important even for moderately large samples. Thus, it can be said that (ii) is often the most practically important of the two, though (i) is certainly theoretically

important. The software here provides diagnostics that assess these two matters, which would be less clear from direct simulation as described later in this section.

Our thinking has been largely influenced by Barndorff-Nielsen (1986, 1991) and Skovgaard (1996, 2001). We note, however, that there has been a parallel, somewhat different and penetrating, thread of development by Donald Fraser and colleagues: Fraser & Reid (1988), Fraser (1991, 2004), Reid (2003) and Brazzale & Davison (2008). The closest point of contact with the particulars of this paper arises in the approximations considered in Section 5. Anthony Davison, starting with Davison (1988), has performed much to promote the ideas reviewed in this paper, particularly for focus on the conditional likelihood function; see Chapter 12 of Davison (2003) and the aforementioned three citations with Davison as co-author.

1.2 Overview of Main Results and Basis for Development

Focus here in these methods is on testing a hypothesis on the value of a smooth scalar function $\psi(\theta)$ of the model parameters. Frequently, it will be best to obtain higher order confidence intervals by testing a grid of hypotheses on ψ roughly spanning a first-order interval of form $\hat{\psi} \pm \mathrm{SE}(\hat{\psi})$, taking the confidence interval as values not 'rejected' by a one-sided test. The package `likelihoodAsy` automates this process. Let $W_\psi(y)$ be the usual generalized likelihood ratio statistic, detailed in (2) of Section 2, with limiting χ^2_1 distribution, and consider one-sided inference based on $r_\psi(y) = \mathrm{sgn}(\hat{\psi} - \psi)\, W_\psi(y)^{1/2}$. Parameters with hats will denote maximum likelihood estimators.Then for observed data y, first-order inference can be based on the result

$$P\left\{r_\psi(Y) \le r_\psi(y); \theta : \psi(\theta) = \psi\right\} = \Phi\{r_\psi(y)\}\left\{1 + O\left(n^{-1/2}\right)\right\},$$

where $\Phi(\cdot)$ is the standard normal distribution function and n is the sample size. The results considered in this paper involve a modification r^*_ψ of this r_ψ, that is commonly denoted by simply r^*, for which the higher order accuracy can be formalised as the second-order result

$$P\left\{r_\psi(Y) \le r_\psi(y); \theta : \psi(\theta) = \psi\right\} = \Phi\{r^*_\psi(y)\}\left\{1 + O\left(n^{-1}\right)\right\}, \tag{1}$$

provided that ψ is within $O(n^{-1/2})$ of its maximum likelihood estimator. There is a similar statement for the upper tail probability. Further aspects of (1) are discussed in the next section in connection with (1*). To clarify an elusive matter, we note that setting the observed value of r^*_ψ equal to 1.96 and solving this equation for ψ provides a lower confidence limit with error probability 0.025; changing the sign provides an upper limit. A result of doing this for both lower and upper confidence limits, for all confidence levels, is shown soon in Figure 2.

In these relations, n will be the number of observations when the dataset consists of independent contributions, otherwise a more general measure such as the Fisher information determinant $|i|$. Note that the error bounds are relative, which is important when the P-values are small. Relations pertaining to (1) are more commonly expressed in terms of asymptotic standard normality of $r^*_\psi(Y)$, but we prefer (1) as being inferentially more clear and direct. Throughout the paper, first, second and third orders refer to powers of $n^{-1/2}$ in expressions such as (1).

The reason for focus on the signed square root of the likelihood ratio is mainly to allow for one-sided tests. Adjustments to the chi-squared LR test $W_\psi(y)$, such as the Bartlett adjustment, do not allow for this. Indeed, it is usual that nominally equi-tailed intervals based on first-order methods have actual error rates differing considerably in the two directions and only average to near the nominal level. The quantity r^* was derived by Barndorff-Nielsen (1986, 1991) in path-breaking work but in a form difficult to compute in general. This is what led to the work by

Pierce & Peters (1992), which however dealt only with exponential families. Various approximations have emerged, and in this paper and the accompanying software, we utilise the version developed by Skovgaard (1996, 2001). The work of Fraser and colleagues referred to previously led to a different version of r^*. Sometimes, workers distinguish notationally between the original r^* and approximations to it; for example Skovgaard uses $\tilde{r}$ for this. Here, we will use r^* to denote any of these, referring to the version to distinguish between approximations. Other approximations to r^* were proposed; Severini (2000, section 7.5), some of which are inferior to that employed here for reasons explained later. We utilise simple simulation, without model fitting, to apply the Skovgaard method but do not consider that as yet another approximation; rather just a way to facilitate broad application of Skovgaard's remarkable advance.

Understanding the basis for (1) emphasises some steps differing from the usual Neyman–Pearson approach, though the end results are nearly the same when the latter arrives at an exactly 'optimal' solution. That optimality obtains largely only for certain inferences in full-rank exponential families and transformation models, beyond which the usual course is to employ the power-maximising principles within the first-order approximation realm. In that case, use of (1) is typically more accurate than approximations ordinarily used. The material sketched in the next two paragraphs comprises the specifics of the modern likelihood-frequentist inference of the title, as indicated at the outset of this section and is further discussed later.

Steps leading to (1) can be thought of in terms of the following:

(i) a highly accurate 'likelihood ratio approximation' to the distribution of the maximum likelihood estimator $\hat{\theta}$,
(ii) a transformation and approximate integration to obtain from that a correspondingly accurate approximation to the distribution of $r_\psi(Y)$ under the hypothesis on $\psi(\theta)$.

The approximation in (i), often called the p^* formula, is novel in the higher order theory. The Jacobian for step (ii) can be difficult to compute, so a main issue is approximating this, here using the Skovgaard approach.

In the simplest development, the likelihood ratio approximation requires that $\hat{\theta}$ be a sufficient statistic, for example see Durbin (1980). In settings where it is not, when the Neyman–Pearson approach usually turns to first-order approximations, the approach outlined here is to condition on an approximate *ancillary statistic* a such that $\hat{\theta}$ is conditionally sufficient—this being very generally applicable. This achieves a sense of 'optimality' differing in principle and sometimes in results, from the power-maximisation of the Neyman–Pearson theory. This emphasis on sufficiency is a key aspect of the approach to inference here, paving the way for second-order approximations. The concepts of ancillarity are that to suitable asymptotic approximation:

(iii) an ancillary statistic a carries information about the *precision* of $\hat{\theta}$, but not the value of θ, that is its distribution is free of θ, and
(iv) $(\hat{\theta}, a)$ is sufficient, and conditionally on a, the estimator $\hat{\theta}$ is sufficient.

This conditioning of (iii–iv) is most important in the considerations of this paper. The ancillarity is almost always approximate. An important version is the Efron–Hinkley ancillary introduced in Section 2, which is essentially the ratio of observed to expected Fisher information, as defined in the next section. When the maximum likelihood estimator is not sufficient, then the observed information varies around its expectation. The reciprocal of the observed information approximates the variance of the maximum likelihood estimator, when conditioning on this Efron–Hinkley approximate ancillary. Another important consequence of such

conditioning is that it renders the maximum likelihood estimator to be a second-order sufficient statistic. This simplifies greatly the matter of finding the ideal inference and in a manner that is actually more effective than the power-maximising Neyman–Pearson theory. This encapsulates what has been termed neo-Fisherian inference; see Pace & Salvan (1997).

We note that use of r^* is also important in full-rank exponential families, where the maximum likelihood estimator is sufficient, and there is no need for ancillary conditioning. In that setting, it is well known, and central to the Neyman–Pearson theory, that when $\psi(\theta)$ is a linear function of the natural parameters, or a ratio of these as in the Student's t-test, the best test whose P-value does not depend on the nuisance parameter is obtained by conditioning on the nuisance parameter estimates. These are referred to as similar tests; see Cox and Hinkley (1974, section 5.2). Pierce & Peters (1992) developed the same r^* as in this paper but for approximating the conditioning for similar tests. This does not require the approximations of Section 5. This will be taken up briefly in Section 6.

1.3 Examples Using `R` *Package*

Before turning to further details of general issues, we offer an example of ancillary conditioning in terms of the `R` software accompanying this paper. The function of the primary routine in this software is to fit the model with and without the hypothesis constraint and then carry out a modest simulation considered in Section 5 for approximating Jacobians to implement the Skovgaard version of r^*. This involves approximating covariances of likelihood quantities by simulation with no model fitting.

Example 1. Weibull regression

Consider a sample of n observations from a Weibull distribution for response times t, including regression-type covariates. The model can be defined in terms of the survival function $S(t_i;\beta,\gamma) = \exp\left[-\left\{t_i^{\gamma}\exp(z_i^{\top}\beta)\right\}\right]$ so that $\theta = (\beta,\gamma)$ where β is a vector of regression parameters for covariates z_i, and the scalar γ governs the 'shape' of the distribution. Inference will be considered not simply for individual coordinates of θ, but for the survival probability or reliability, at a given time t_0, and for a specified covariate vector z_0. The interest parameter ψ will be represented as the log reliability $\psi(\beta,\gamma) = -t_0^{\gamma}\exp(z_0^{\top}\beta)$, where using this logarithmic representation does not affect r^*_{ψ}, but it can affect the numerical behaviour of the constrained maximisation routine.

For the Weibull model, the maximum likelihood estimator $\hat{\theta}$ is not a sufficient statistic. The logarithms of the response times follow a location-scale model, with a regression-type form for the location parameter. For such models, there is a well-known exact ancillary, of dimension $n - \dim(\theta)$ referred to as the [spacing] configuration of the sample; see Lawless (1973, 2003, appendix E) and Davison (2003, example 5.21). Though exact inference conditional on this ancillary can be accomplished with multi-dimensional numerical integration, this is seldom used in practice. The methods here approximate well that conditional inference, even though they are based on conditioning on a more general approximate ancillary, along lines considered in the following section.

We employ the commonly used data from Feigl and Zelen (1965, table 1 left panel) with $n = 17$ and $\dim(\beta) = 2$, involving simple linear regression on `log(WBC)` of the log failure rate for leukaemia survival. We will choose for defining our survival probability interest parameter values t_0 and z_0 such that, for our dataset, the maximum likelihood estimate of the reliability is 0.10. For a hypothesis on this with P-values small enough to be of interest in comparing first-order and second-order inferences, we will test that the reliability is 0.03, which is approximately a lower 97.5% confidence limit based on first-order Wald test methods.

```
loglik.Wbl <- function(theta, data)
{
  logy <- log(data$y)
  X <- data$X
  loggam <- theta[1]
  beta <- theta[-1]
  gam <- exp(loggam)
  H <- exp(gam * logy + X %*% beta)
  out <- sum(X %*% beta + loggam + (gam - 1) * logy - H)
  return(out)
}

gendat.Wbl <- function(theta, data)
{
  X <- data$X
  n <- nrow(X)
  beta <- theta[-1]
  gam <- exp(theta[1])
  data$y <- (rexp(n) / exp(X %*% beta)) ^ (1 / gam)
  return(data)
}

psifcn.Wbl <- function(theta)
{
  beta <- theta[-1]
  gam <- exp(theta[1])
  y0 <- 130
  x0 <- 4
  psi <- -(y0 ^ gam) * exp(beta[1] + x0 * beta[2])
  return(psi)
}
```

Figure 1. *Functions provided by user for Weibull regression example.*

The functions to be provided by the user of our software are shown in Figure 1. The function `psifcn.Wbl` defines the interest function, and the other functions pertain to the case of no censoring. We note that although allowing for censored data as considered below entails only a minor change in the likelihood routine, the data-generation routine will then need to involve a probability model for the censoring.

For the example dataset, the main routine `rstar` in `likelihoodAsy` then returns, for testing $\psi = \log(0.03)$,

$$\begin{aligned} r_\psi &= 1.67 \quad (P = 0.048), \\ r_\psi^* &= 2.10 \quad (P = 0.018), \\ \text{Wald} &= 1.96 \quad (P = 0.025), \end{aligned}$$

where the latter is the Wald statistic in representation $(\hat{\psi} - \psi)/\text{SE}(\hat{\psi})$. Confidence limits are shown in Figure 2. The displayed P-values have no general inferential relevance beyond providing some points on the curves in Figure 2 regarding confidence limits. It was mentioned

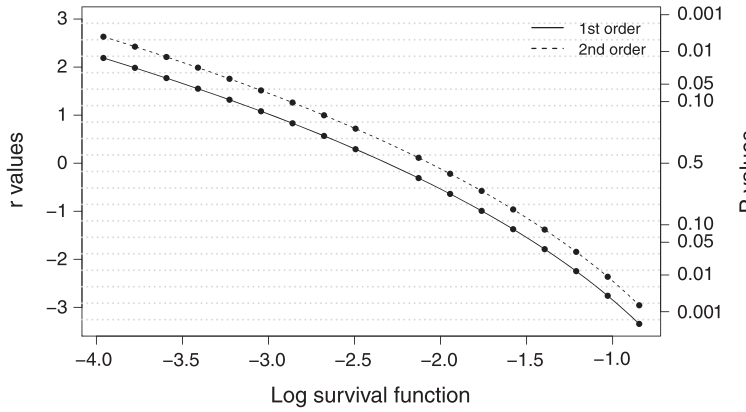

Figure 2. *Upper and lower one-sided confidence limits at all levels, using first-order and second-order likelihood ratio methods. The* $\boldsymbol{P}$*-values are the one-sided error rates given by* $1 - \Phi(|r^*|)$.

earlier that our approach provides diagnostic information on the shortcomings of first-order inferences. This is detailed later, but we can say now that about 63% of the adjustment $r^* - r$ is due to presence of the 2 nuisance parameters, with the remainder being due to the specifics of limited information with only 17 observations.

These results are not atypical, for settings with few nuisance parameters; with more nuisance parameters, the higher order adjustment is often much larger. As considered at the end of Section 2, we can evaluate the accuracy of the fundamental approximation (1) by simulating the distribution of $r_\psi(Y)$, using Weibull datasets with parameter values fitted to the analysis dataset under the hypothesis on ψ. This is a parametric bootstrap approach to hypothesis testing, to be discussed in later sections. The result with 50 000 simulation trials is that, empirically, $P\left\{r_\psi(Y) > r_\psi(y); \theta : \psi = \log(0.03)\right\} \doteq 0.021$, which compares favourably to $1 - \Phi(r_\psi^*) = 0.018$. Though the Wald statistic is here slightly more accurate than the unadjusted likelihood ratio test, one should not think this is typical. The problem is that the Wald test is sensitive to the choice of parametrization, and it is not always easy to find a good choice for this.

It is not difficult to allow with the package for censoring in such analyses. This involves specifying a censoring model in terms of the `gendat.Wbl` function. In principle, censoring models, if applicable, must be specified for inferences going beyond first order; results depend on more than the likelihood function. To exemplify this very simply, we have carried out the inference based on modifying these data, under a censoring model where the largest 5 failure times are censored at the just-preceding failure time. The code for this is given in the vignette of the documentation for `likelihoodAsy` package. For that, we alter the hypothesis to be the 0.975 lower Wald confidence limit for the censored data, as performed for the example previously, which is now −2.07. Results become

$$\begin{aligned} r_\psi &= 1.59 \quad (P = 0.056), \\ r_\psi^* &= 1.99 \quad (P = 0.023), \\ \text{Wald} &= 1.96 \quad (P = 0.025). \end{aligned}$$

We now briefly consider another example emphasising that the adjustments can make a practical difference in real data settings, even when the sample size is not particularly small, when there are several nuisance parameters.

Example 2. Veterans Administration Cancer Data

These data are in the dataset veteran of the `R` package `survival`, which is also given in an appendix of Kalbfleisch & Prentice (2002). Results here are for the selection with cell type squamous comprising 35 observations of which 4 are censored. Analysis is of survival time and there are five covariables: `treatment indicator`, `Karnofski performance`, `diagnosis time`, `age` and `prior therapy`. For a Weibull model analysis, the user-defined functions can be similar to those for Example 1. In the data selection used, the few individuals are censored at apparently random times, rather than at termination of the study as we investigated for Example 1. We use here a censoring model with exponentially distributed censoring times, calibrated to achieve about the observed amount of censoring. For comparison of first-order and second-order methods, we test that the age coefficient is equal to the 97.5% upper Wald confidence limit 0.074. The P-value based on r is 0.031 and that based on r^* is 0.011. About 90% of the inadequacy of the first-order result is due fitting the considerable number of nuisance parameters, with the remainder due to limited information with only the 35 observations.

2 General Preliminary Issues

Let $p(y;\theta)$ be the density (or probability mass function) for a dataset Y that is not necessarily a collection of independent observations, with $\dim(\theta) \geq 1$. The observations can be continuous or discrete, and the primary regularity condition is that $p(y;\theta)$ is a differentiable function of the parameter, ruling out typical settings where the set of y-values where $p(y;\theta) > 0$ depends abruptly on θ. The likelihood function $L(\theta; y)$ is any function that is proportional to $p(y;\theta)$ for given y, and we write $\ell(\theta; y)$ for the log likelihood function. The *observed information* matrix for observed data y is $j(\theta) = -\partial^2\ell(\theta; y)/\partial\theta\partial\theta^\top$, $\hat\theta$ is the maximum likelihood estimator of θ, and we write $\hat\jmath$ for the matrix $j(\hat\theta)$. When $\hat\theta$ is not a sufficient statistic, $\hat\jmath$ varies around the expected information $\hat\imath = E\{j(\theta)\}_{\theta=\hat\theta}$, suggesting that the inference should reflect some samples being more informative than others—an issue central to this paper.

Inference is here developed for any given smooth scalar function of the parameter $\psi = \psi(\theta)$, which is referred to as the *interest parameter*. It is convenient to utilise a $p-1$ dimensional *nuisance parameter* λ such that the transformation $\theta \leftrightarrow (\psi, \lambda)$ is 1-1, but it is important that results be independent of the arbitrary representation of the nuisance parameter. Note that in `likelihoodAsy`, the user does not need to specify a form for the nuisance parameter; one is employed in the analysis program, but it is determined in the code for `rstar`. We will first assume that the distribution of y is continuous and then deal in Section 6 with discrete settings.

The signed square root likelihood ratio statistic is defined as

$$r_\psi(y) = \mathrm{sgn}(\hat\psi - \psi)\sqrt{2\left\{\ell(\hat\theta; y) - \ell(\tilde\theta; y)\right\}}, \tag{2}$$

where $\tilde\theta = (\psi, \hat\lambda_\psi)$ is the constrained maximum likelihood estimator. Note that (2) does not depend on the specific representation of the nuisance parameter. We will throughout often use the hat and tilde to denote the unconstrained and constrained estimators. As noted, this is a signed square root of the usual χ^2_1 statistic, and as described in Section 1, the random variable $r_\psi(Y)$ has to first order $O_p(n^{-1/2})$ a standard normal distribution under the hypothesis. The aim is to improve on this approximation to second order as in (1).

In the following sections, we define the quantity r^*_ψ used in (1). One of the key issues is a suitable approximate ancillary statistic as indicated Section 1. Any locally second-order ancillary will meet the needs for this paper. The main result in (1) and (1*) in the next discussion

will then be unique to second order. However, we believe that basic ancillarity issues will be more readily understood, in terms of the following specific choice, and that this may avoid some misconceptions about what we have in mind. For our needs, the ancillary information can be transparently, generally and effectively based on the ratio of observed to expected information that is the matrix $\hat{\imath}^{-1}\,\hat{\jmath}$. In order for the distribution of the ancillary to be more nearly constant in θ, we may rescale $\hat{\imath}^{-1}\,\hat{\jmath}$ by dividing it, in a matrix sense, by $n^{1/2}$ times its estimated asymptotic standard deviation, resulting in an ancillary statistic $a = \Gamma(\hat{\theta})\,\hat{\imath}^{-1}\,\hat{\jmath}$, where the precise form of $\Gamma(\hat{\theta})$ is implicitly given in Skovgaard (1985, eq. (2.1)); see Endnote 1. This ancillary, reflecting ideas of Fisher and others, is called the Efron–Hinkley ancillary, studied in the paper of Efron & Hinkley (1978) that is notable for its exposition.

Skovgaard (1985) established that this Efron–Hinkley ancillary is locally second-order ancillary, meeting needs for this paper. The meaning of this is that the distribution of a depends on θ only in terms of $O(n^{-1})$ for θ-variations of $O(n^{-1/2})$, and in this same sense, $(\hat{\theta}, a)$ is approximately sufficient.

Important aspects of modern likelihood asymptotics involve what we refer to as protection for large deviations. Normal deviations refer to regions where $\|\hat{\theta}-\theta\| = O_p(n^{-1/2})$, that is the usual inferential region of interest. In that region, the main results in this paper are of second order $O(n^{-1})$. Large deviations refer to regions where $\|\hat{\theta}-\theta\| = O_p(1)$, where P-values will typically be extremely small. The 'protection' arises from employing derivations such that, even when P-values are very small, the approximations maintain relative error $O(n^{-1/2})$. It is not that great accuracy is desired for testing hypotheses far from estimate, but rather that imposing this approach leads to methods, which do not deteriorate as θ moves away from $\hat{\theta}$ and thus are particularly accurate for normal deviations.

We now consider (1) with more detail regarding the ancillary conditioning. Let a be a locally second-order ancillary such as the Efron–Hinkley choice. Then for testing an hypothesis on $\psi = \psi(\theta)$,

$$P\left\{r_\psi(Y) \le r_\psi(y)|a;\psi,\lambda\right\} = \Phi\{r^*_\psi(y)\}\left\{1 + O(|\hat{\psi}-\psi| \times \|a\|)\right\}, \qquad (1^*)$$

where a is expressed so that it is $\|a\| = O(n^{-1/2})$. If both the parameter and ancillary deviations are of normal type, then the relative error becomes $O(n^{-1})$. This result remains valid without the ancillary conditioning, as written in (1), as it is a key feature of likelihood ratio statistics that they are to second order independent of any ancillary; see McCullagh (1984), Severini (1990) and Severini (2000, section 6.4.4). Note the implication in (1*) that such P-values do not, to second order, depend on the nuisance parameter. If the deviation regarding ψ is large, then the error bound is $O(n^{-1/2})$, and for some second-order approximations, the measure of error is not that small; see Endnote 2. Skovgaard (1996, 2001) proved these claims, for the version of r^* we employ here.

Because (1*) pertains to the true distribution of $r_\psi(Y)$, an alternative approach to computing P-values would involve direct simulation, that is the 'parametric bootstrap', with references to follow. An important issue is that a standard unconditional simulation leads to inferences agreeing with the ancillary conditioning in (1*), because the likelihood ratio statistic is to second order independent of any ancillary; see Davison *et al.* (2003), DiCiccio *et al.* (2001), DiCiccio *et al.* (2015) and Young (2009). This asymptotic independence also obtains with other pivotals agreeing sufficiently well with the likelihood ratio, such as the Wald statistic using observed information. It is well known and considered in the references just given that if the constrained maximum likelihood estimator $\tilde{\theta}$ is used for such simulation, then as the number of trials approaches infinity, the results approximate to $O(n^{-3/2})$ quantiles of the distribution of $r_\psi(Y)$.

It is attractive to many to employ the parametric bootstrap as an alternative to use of r^*, on the grounds that it is more transparent and the results agree well. There are however reasons to

prefer the use of r^*. It was shown by DiCiccio & Young (2008) that it can require close to 50 000 bootstrap trials to achieve comparable accuracy. More importantly, fitting models under the hypothesis is often computationally challenging, particularly when ψ is not simply a coordinate of θ, and this fitting becomes even more problematic when it must be carried out a large number of times. Evolution of the package `likelihoodAsy` has been attentive to this problem, for which long-term developments in numerical optimisation, such as the augmented Lagrangian method (e.g. Nocedal & Wright, 2006), provide an effective resolution. Note that, however, regardless of the optmization method employed, a recommended strategy is to start from the unconstrained estimate and work toward the constrained solution. Because of this matter, it is usually simpler to apply the r^* function in that package than to carry out a parametric bootstrap using other software for the constrained optimisation. Further, byproducts of r^* not provided by the parametric bootstrap are diagnostics INF and NP regarding the source and magnitude of improvement with second-order approximations. These are raised in Section 4 at Eq. (13).

3 The Likelihood Ratio Approximation to the Density of $\hat{\theta}$

The argument here follows Durbin (1980), which clarifies a key aspect of modern likelihood theory with a rich history; see Endnote 3. What we call the likelihood ratio approximation is often called Barndorff-Nielsen's p^* formula (Barndorff-Nielsen, 1983; Barndorff-Nielsen & Cox, 1994). The argument as summarised here is in detail heuristic, and we comment on that afterwards.

The likelihood ratio approximation to the density of $\hat{\theta}$, when this is a sufficient statistic of $\dim(\theta) \geq 1$, and hence its distributions belong to a full-rank exponential family, is

$$\begin{aligned} p^*(\hat{\theta};\theta) &= \frac{|j(\hat{\theta})|^{1/2}}{(2\pi)^{p/2}} \frac{p(y;\theta)}{p(y;\hat{\theta})} \\ &= p(\hat{\theta};\theta)\left\{1 + O(n^{-1})\right\}, \end{aligned} \tag{3}$$

where $j(\theta)$ is the observed Fisher information, which in this case is also the expected information. As for other approximations in this paper, the error specified in (3) is for $\|\hat{\theta} - \theta\| = O(n^{-1/2})$ and is otherwise $O(n^{-1/2})$. To derive this, consider the following identities, noting that due to sufficiency, the ratio $p(y|\hat{\theta};\theta)/p(y|\hat{\theta};\hat{\theta})$ is unity,

$$p(\hat{\theta};\theta) = \frac{p(\hat{\theta};\theta)}{p(\hat{\theta};\hat{\theta})}\, p(\hat{\theta};\hat{\theta}) = \frac{p(y|\hat{\theta};\theta)}{p(y|\hat{\theta};\hat{\theta})} \times \frac{p(\hat{\theta};\theta)}{p(\hat{\theta};\hat{\theta})}\, p(\hat{\theta};\hat{\theta}) = \frac{p(y;\theta)}{p(y;\hat{\theta})}\, p(\hat{\theta};\hat{\theta}). \tag{4}$$

We now assume that $p(\hat{\theta};\theta)$ admits an Edgeworth expansion, with conditions for this being given by Durbin (1980). When this is evaluated at $\theta = \hat{\theta}$, the correction term to the base Gaussian density vanishes (see remark in penultimate paragraph of this section), and for that Gaussian term, the exponent is zero so that to second order $p(\hat{\theta};\hat{\theta}) = (2\pi)^{-p/2}\,|j(\hat{\theta})|^{1/2}$, which provides (3). This $p^*(\hat{\theta};\theta)$ does not ordinarily integrate exactly to unity, and the accuracy is improved by one power of $n^{1/2}$ by normalising it, but this is not employed for our needs.

It is the key to this result, and in a more general sense, to much of modern likelihood asymptotics, that in the only approximation made here, that to $p(\hat{\theta};\hat{\theta})$, the true parameter 'tracks' the estimator $\hat{\theta}$ so that the resulting approximation to $p(\hat{\theta};\theta)$ is good not only for $\hat{\theta}$ near θ but also for large deviations of $\hat{\theta} - \theta$, as mentioned in connection with (1*).

For the case that $\hat{\theta}$ is not sufficient, we proceed somewhat more heuristically. It is a central aspect of higher order likelihood asymptotics that, to suitable approximation, there is an ancillary statistic a such that the quantity $(\hat{\theta}, a)$ is sufficient, and $\hat{\theta}$ is sufficient in the model for the conditional distribution of $\hat{\theta}|a$. These conditions mean that to a related order of approximation, the distribution of a does not depend on θ. In addition to showing that the Efron–Hinkley statistic is locally second-order ancillary in that sense, Skovgaard (1985) provided 'information loss' results in the direction of the conditional sufficiency just considered. More fully, Reid (1988, section 3) notes that several researchers, primarily Barndorff-Nielsen and McCullagh, had already considered the second-order conditional sufficiency in the model for $\hat{\theta}|a$. The second-order results of this section hold for any choice of first-order ancillary; see Pace and Salvan (1997, section 11.2), which implies that it is locally second-order ancillary as considered by Cox (1980) and Skovgaard (1985). In a slightly different statement, Reid (1988, section 3) notes that the results in this section hold for any second-order ancillary.

Because of these approximate sufficiency considerations, the same argument as mentioned previously applies conditionally, leading now to the same approximation formula, but interpreted as approximating the density conditional on an ancillary,

$$\begin{aligned} p^*(\hat{\theta}|a;\theta) &= \frac{|j(\hat{\theta};\hat{\theta},a)|^{1/2}}{(2\pi)^{p/2}} \frac{p(y;\theta)}{p(y;\hat{\theta})} \\ &= p(\hat{\theta}|a;\theta)\left\{1 + O(n^{-1})\right\} . \end{aligned} \tag{5}$$

Note that in the argument (5), the omitted term $p(y|\hat{\theta},a;\theta)/p(y|\hat{\theta},a;\hat{\theta})$ is no longer exactly unity but is $1 + O(n^{-1})$ because of the second-order sufficiency of $(\hat{\theta}, a)$. Though the observed and expected information were identical in the sufficiency setting of (3), they are no longer so, and it is more accurate to use the observed information in (5), as the appropriate variance for the Edgeworth expansion. The most remarkable aspect of this approximation is that the formula is the same as when $\hat{\theta}$ is sufficient, with the understanding that the 'observed information' in (3) actually coincided with the expected information, which is not the case in (5). Nevertheless, the generality of this may be largely the reason that (5) is often referred to as Barndorff-Nielsen's '*magic formula*', e.g. Efron (1998).

It is known that (3), after rescaling to integrate to unity, is exact for location-scale and all other transformation models. This was realised by Fisher (1934), long before the other developments of this section began. In the same paper, Fisher suggested that $\hat{\theta}$ supplemented by the first few derivatives of the log likelihood at $\hat{\theta}$ would be asymptotically sufficient.

Reasons to consider the previous discussion as heuristic include the need to deal with asymptotically negligible bias in $\hat{\theta}$, complicating the treatment of $p(\hat{\theta};\hat{\theta})$ in the argument, the matter of existence of the Edgeworth expansion in the general setting, and many matters glossed over in deriving (5) when $\hat{\theta}$ is not sufficient. These issues involving $p(\hat{\theta};\hat{\theta})$ apply whether or not $\hat{\theta}$ is sufficient, and for the sufficient case were considered by Durbin (1980). Having raised these matters of heuristic reasoning, we add that it seems remarkably difficult to obtain a rigorous proof to the desired result in its full generality; see for example Reid (1988), section 2.2.

The approximations (3) and (5) are of limited practical value for direct use, because as $\hat{\theta}$ varies, one must correspondingly vary y, keeping fixed the ancillary in the case of (5). In this respect, the final factor in (5) will depend to second order on the choice of ancillary. Although for given data y, (5) is simply the likelihood function, what is being approximated is not an object involving given data but the density as a function of $\hat{\theta}$. Thus, it might be said that (5) is

deceptively simple, although the text by Butler (2007) gives useful and interesting applications. The main point, to follow, is that for approximating the distribution of r_ψ, the p^* formula becomes far more useful.

4 Corresponding Distribution of r_ψ

We first consider this when $\dim(\theta) = 1$. The distribution of r_θ derived from the likelihood ratio approximation to the distribution of $\hat{\theta}$ has density, under regularity conditions mainly involving monotonicity,

$$p^*(r_\theta|a;\theta) = |\partial r_\theta/\partial\hat{\theta}|^{-1}\, p^*(\hat{\theta}|a;\theta)\,. \tag{6}$$

This is not convenient to use, and we make a further second-order approximations as follows. Note that $\partial r_\theta/\partial\hat{\theta} = [\partial\{\ell(\hat{\theta};\hat{\theta},a)-\ell(\theta;\hat{\theta},a)\}/\partial\hat{\theta}]/r_\theta$, from differentiating r_θ^2 in definition (2). This Jacobian term indicates why *sample space derivatives*, that is likelihood partial derivatives with respect to $\hat{\theta}$, are central to higher order likelihood asymptotics. When $\hat{\theta}$ is sufficient, no ancillary arises, and the sample space derivatives may be straightforward. However, when an ancillary must be held fixed, this is seldom tractable. That is, it would often be nearly impossible to express the likelihood in terms of $(\hat{\theta},a)$ in order to carry out the partial differentiation, particularly when the ancillary is only approximate. Our resolution of this is to leave the theoretical statement in such possibly intractable terms but for implementation to employ a general method for approximating these sample space derivatives, which is detailed in Section 5.

Denote by u_θ the sample space derivative $\partial\{\ell(\hat{\theta};\hat{\theta},a)-\ell(\theta;\hat{\theta},a)\}/\partial\hat{\theta}$ divided by $|j(\hat{\theta})|^{1/2}$, so (6) can be expressed as

$$p^*(r_\theta|a;\theta) = |u_\theta/r_\theta|^{-1}\,(2\pi)^{-1/2}\,\exp(-r_\theta^2/2)\,. \tag{7}$$

This means that to second-order approximation, u_θ must be a function of r_θ when holding fixed a, and Barndorff-Nielsen (1986, section 3.2) gave results showing that to second order u_θ/r_θ is quadratic in r_θ with coefficients that are data-independent functions of θ. It is fundamental to the nature of our aims that $r_\theta^* = r_\theta + O_p(n^{-1/2})$, and we note that $u_\theta/r_\theta = 1 + O_p(n^{-1/2})$ for $|\hat{\theta}-\theta| = O_p(n^{-1/2})$. Sweeting (1995) analysed such structures as (7) under these conditions, referring to such densities as 'near-normal'. Through raising $|u_\theta/r_\theta|^{-1}$ to the exponential, completing the square, and dropping the term $(u_\theta/r_\theta)^2$, we have

$$p^*(r_\theta|a;\theta) = (2\pi)^{-1/2}\,\exp\{-(r_\theta^*)^2/2\}\,, \tag{8}$$

where

$$r_\theta^* = r_\theta + r_\theta^{-1}\,\log(u_\theta/r_\theta)\,. \tag{9}$$

It also follows from results in Barndorff-Nielsen (1986, section 3.2) that, to second order, r_θ^* is monotonically increasing in r_θ, so one can compute tail probabilities in terms of these pivotals, and thus, we have formulations as introduced at the outset in (1),

$$P\,\{r_\theta(Y) \le r_\theta(y)|a;\theta\} = \Phi\{r_\theta^*(y)\}\,\left\{1 + O\left(n^{-1}\right)\right\}\,. \tag{10}$$

It is somewhat more common to consider the higher order distribution of r^*, as noted in Section 1 and discussed in Endnote 2. The result (9) as stated holds under the normal deviation condition $|\hat{\theta}-\theta| = O_p(n^{-1/2})$, which is required for the likelihood ratio approximation to the

distribution of $\hat{\theta}$ and other steps following (7). See Endnote 2 regarding second-order versus third-order approximations and also remarks following (1*).

We now turn to the case where $\dim(\theta) > 1$, expressing $\theta = (\psi, \lambda)$ as in Section 2. The material is intricate and could be only skimmed on a first reading; see Endnote 4. Recall that the likelihood ratio approximation to the density of $\hat{\theta}|a$ applies in the case that $\dim(\theta) > 1$. Thus, the changes from the argument previously are that in transforming from $p^*(\hat{\theta}|a;\theta)$ to the distribution of r_ψ, the Jacobian is for the p dimensional transformation from $\hat{\theta}$ to $(r_\psi, \hat{\lambda}_\psi)$, and we must further integrate out $\hat{\lambda}_\psi$ to obtain the distribution of r_ψ. The standard approach for results in this section is to express the approximate marginal distribution of r_ψ in the form

$$p(r_\psi|a;\theta) = \frac{p(r_\psi, \hat{\lambda}_\psi|a;\theta)}{p(\hat{\lambda}_\psi|r_\psi, a;\theta)} \tag{11}$$

and employ the likelihood ratio approximation to the numerator and denominator. The numerator involves, similarly to in (6), a Jacobian that now becomes $|\partial(r_\psi, \hat{\lambda}_\psi)/\partial\hat{\theta}|^{-1}$, and a further sample space derivative that can be expressed as

$$\frac{\partial^2 \ell(\psi, \hat{\lambda}_\psi; \hat{\theta}, a)}{\partial\lambda\partial\hat{\lambda}^\top} = \left.\frac{\partial^2 \ell(\psi, \lambda; \hat{\theta}, a)}{\partial\lambda\partial\hat{\lambda}^\top}\right|_{\lambda=\hat{\lambda}_\psi} .$$

The sample space derivative raised for (7) now becomes $\partial\{\ell_P(\hat{\psi}; \hat{\theta}, a) - \ell_P(\psi; \hat{\theta}, a)\}/\partial\hat{\psi}$, where $\ell_P(\cdot)$ denotes the profile log likelihood, that is $\ell_P(\psi; \hat{\theta}, a) = \ell(\psi, \hat{\lambda}_\psi; \hat{\theta}, a)$. The likelihood ratio approximation to the denominator is straightforward, upon observing that the statistic (r_ψ, a), as opposed to simply a, is a suitable ancillary for the smaller family, where ψ is considered as fixed.

The details of obtaining the approximation $p^*(r_\psi|a;\theta)$ in this manner are given in section 7.4 of Severini (2000), up through his (7.4) for the near-normal distribution of r_ψ, which can then be dealt with in the manner of steps between our (7) and (9); see also Barndorff-Nielsen & Cox (1994), section 6.6. The result is again our formula (8), but with a more general definition of r^*_ψ, that can be expressed as

$$r^*_\psi = r_\psi + r_\psi^{-1} \log(C_\psi^{-1}) + r_\psi^{-1} \log(\tilde{u}_\psi / r_\psi) , \tag{12}$$

where

$$C_\psi^{-1} = \left|\frac{\partial^2 \ell(\psi, \hat{\lambda}_\psi; \hat{\theta}, a)}{\partial\lambda\partial\hat{\lambda}^\top}\right| \{|\hat{j}_{\lambda\lambda}|\,|\tilde{j}_{\lambda\lambda}|\}^{-1/2} , \tilde{u}_\psi = \left[\frac{\partial\{\ell_P(\hat{\psi}; \hat{\theta}, a) - \ell_P(\psi; \hat{\theta}, a)\}}{\partial\hat{\psi}}\right] \hat{J}^{-1/2}_{\psi\psi|\lambda} .$$

Here, recall that the tilde denotes evaluation at $(\psi, \hat{\lambda}_\psi)$, $\tilde{u}_\psi$ is given the sign of r_ψ and $j_{\psi\psi|\lambda}$ denotes the adjusted information for ψ, as defined shortly. The two final terms in (12) derive almost entirely from the Jacobian re-expressed in likelihood terms and arranged into two parts for reasons to follow. The final one of these two terms of (12) is essentially the same as in (9), except for being defined in terms of the profile likelihood. The intricacy of the aforementioned relations is discussed in Endnote 4. The penultimate term is a new object corresponding to Barndorff-Nielsen's modified profile likelihood (MPL) $L_{MP}(\psi; y) \propto L_P(\psi; y)\, C_\psi$ for the setting of a scalar parameter ψ. This modified likelihood pertains specifically to allowing for effects of fitting nuisance parameters λ. The MPL, in contrast to r^*, applies to the case $\dim(\psi) > 1$. Higher order test statistics for this case are proposed in Skovgaard (2001) and Fraser *et al.* (2016).

Related to MPL is the Cox–Reid approximate conditional likelihood (Cox & Reid, 1987), given by omitting the troublesome term $|\partial^2 \ell(\psi, \hat{\lambda}_\psi; \hat{\theta}, a)/\partial\lambda\partial\hat{\lambda}^\top|$ in C_ψ^{-1}. However, as those authors noted, this results in loss of the invariance to the choice of representation of the nuisance parameter. They dealt with that alternatively by requiring that the parameters (ψ, λ) be 'orthogonal' to each other; however, there is not a unique choice in this respect.

It is useful and common to express (12) as

$$r_\psi^* = r_\psi + \mathrm{NP}_\psi + \mathrm{INF}_\psi \,, \tag{13}$$

referring to the terms as the nuisance parameter and information adjustments (Pierce & Peters, 1992; Barndorff-Nielsen & Cox, 1994, section 6.6.4). Neither of these adjustments depends on the representation of the nuisance parameter. The NP adjustment can be substantial when $\dim(\lambda)$ is not small, even for moderately large samples. Whether this occurs depends on the structure of the model, in ways difficult to ascertain without computing at least an approximation to NP. The INF adjustment is often small unless the data are so limited that the ψ inference is of marginal practical value; that is the adjusted information $j_{\psi\psi|\lambda} = j_{\psi\psi} - j_{\psi\lambda}\, j_{\lambda\lambda}^{-1}\, j_{\lambda\psi}$ is quite small. Both terms of the decomposition are conceptually important, but only the NP adjustment is commonly important in practice.

It is easily seen that the MPL introduced just previously can be expressed exactly as $L_{MP}(\psi; y) \propto L_P(\psi; y)\, \exp(-r_\psi\, \mathrm{NP}_\psi)$, where L_P denotes the profile likelihood. This is only for the case that $\dim(\psi) = 1$, and note for this case that the MPL function does not involve the INF adjustment. It is, however, true that the NP adjustment is usually much larger than INF. The INF adjustment mainly captures the nature of the model that is not reflected in the likelihood function, for example the distinction between binomial and inverse binomial sampling (Pierce & Bellio, 2006). Since that distinction is not included in Bayesian inference, the MPL is more suitable for a prior involving only the interest parameter in that setting. For frequentist inference, though, it is hard to justify using the modified profile likelihood as opposed to inferential summaries as that provided by Figure 2.

Pierce & Bellio (2006) have used the decomposition (13) to investigate extra-likelihood aspects of higher order inference in regard to censoring and stopping rules. They found that in some generality, the choice of stopping rule has $O(n^{-1})$ effect on NP but no smaller than $O(n^{-1/2})$ effect on INF. On the other hand, in some generality, the choice of censoring model has $O(n^{-1})$ effect on both NP and INF.

We close this section with two further examples; the first of which involves non-independent observations, and the second illustrates using numerical integration to evaluate the likelihood function.

Example 3. Autoregression model of order 1

We consider use of the `R` package for inference about the correlation of an AR(1) process, with the mean and dispersion of the process as a nuisance parameters. The model is

$$y_i - \mu = \rho\,(y_{i-1} - \mu) + \varepsilon_i \,, i = 1, \ldots, n \,,$$

with independent Gaussian errors satisfying $\mathrm{var}(\varepsilon_i) = \sigma^2$. The full parameter is $\theta = (\mu, \sigma, \rho)$, and we will mainly consider $\psi(\theta) = \rho$. The functions to be provided by the user for inference about both the mean and the correlation are in Figure 3, where `Gamma1` is the inverse of the autocorrelation matrix. In that code, the parameter ρ is reparametrized as Fisher's z transformation. Note that the inverse of this transformation is performed in `psifun.rho` rather than in `likAR1`, because the optimisation is performed on the `theta` scale. Even though our approach is in principle unaffected by choice of parametrization, it makes the optimisation

go more smoothly if constrained ranges, such as $-1 < \rho < 1$, are mapped into $(-\infty, \infty)$. Similarly in general, it can be important to avoid non-negativity constraints by a log transform.

Higher order likelihood inference for this example was considered by Lozada-Can & Davison (2010). An interesting aspect of this is that in order to apply the Fraser approach mentioned earlier, they needed to utilise a special ancillary based on a martingale representation of the AR(1) process, in contrast to the usual ones for independent observations, or the general ancillary based on $\hat{\imath}^{-1}\,\hat{\jmath}$ that we have in mind here. For their dataset of 48 observations of luteinising hormone levels measured at 10-min intervals given in the `R` package `MASS`, they invert the hypothesis testing at 95% level to obtain r^*-based confidence limits for the mean of the process. With our methodology, such results agree with theirs to the three digits reported in their paper. However, as expected for this sample size, with few nuisance parameters, the confidence limits based on r will often agree closely with those based on r^*.

Thus, we consider inference about the correlation of the process, which is somewhat more challenging. For this rather large sample size of $= 48$ inferences from first-order and higher order methods in the lower part of a confidence interval are quite similar. However, there is more distinction in the upper part of a confidence interval, where a first-order 95% one-sided confidence limit for ρ equal to 0.765 based on r is found to be an 88% limit based on r^*. The maximum likelihood estimate of ρ is 0.574 ± 0.116. For testing $\rho = 0.765$, we find $r_\psi = -1.643$ (P=0.050) and $r^*_\psi = -1.155$ (P=0.124). The NP and INF adjustments are 0.36 and 0.13. This r^*-based P-value was confirmed by simulation of 50 000 trials using the parametric bootstrap method discussed at the end of Section 2, yielding $P = 0.131$. For large values of ρ, inference strongly depends on ancillary conditioning, such as on the ratio of observed to expected information; see Johansen (1995).

Example 4. Binomial overdispersion

We now consider one of the standard models for overdispersion in binomial data; namely that $\log\{p_i/(1-p_i)\} = z_i^\top \beta + u_i$, where the u_i are independent $N(0, \sigma^2)$ random variables. This form of modelling, now widely used in more general settings as seen in McCulloch *et al.* (2008), was first considered by Pierce and Sands (TR no. 46, 1975, *Extra-Bernoulli Variation in Binary Data*), and we use a dataset from the text by Finney (1947) that they took as a motivating example.

We note that when numerical differentiation of the log likelihood is to be used, it is important that numerical integration of random effects be highly accurate. This led us to using the Gauss–Hermite quadrature of the code given here, as indicated by the `gq` object that is included in the dataset. In particular, the `gq` object includes the quadrature nodes and weights that have to be at least around 80 to achieve a good accuracy in the r^* computation. Further details are included in the vignette of the documentation of the `likelihoodAsy` package.

The Finney data comprises 10 observations considered to be binomial with numbers of trials about 30 and a single covariable that is the 'dose' for the bioassay. The estimated logit slope is 1.44 ± 0.18 but that standard error is suspect because the residual deviance is 36.25 on 8 d.f., presenting evidence for large binomial overdispersion. Finney's proposed resolution is to multiply the parameter estimate covariance matrix from the binomial analysis by the mean residual deviance 36.25/8 = 4.5. This increases the estimated standard error of the slope estimate from 0.18 to 0.38. The implicit rationale for this assumed that the excess variance over binomial is proportional to the binomial variance, roughly $p\,(1-p)$. The model we employ, and is now widely accepted, has a variance function different from this, with the excess variance being approximately proportional to $\{p\,(1-p)\}^{-1}$. This distinction is studied in detail in the Pierce and Sands TR cited previously.

```
likAR1 <- function(theta, data)
{
  y <- data$y
  mu <- theta[1]
  sigma2 <- exp(theta[2] * 2)
  z <- theta[3]
  rho <- (exp(2 * z) - 1) / (1 + exp(2 * z))
  n <- length(y)
  Gamma1 <- diag(1 + c(0, rep(rho^2, n-2), 0))
  for(i in 2:n)
    Gamma1[i,i-1]<- Gamma1[i-1,i] <- -rho
  lik <- -n/2 * log(sigma2) + 0.5 * log(1 - rho^2) - 1 / (2 * sigma2) *
          mahalanobis(y, rep(mu,n), Gamma1, inverted = TRUE)
  return(lik)
}

genDataAR1 <- function(theta, data)
{
  out <- data
  mu <- theta[1]
  sigma <- exp(theta[2])
  z <- theta[3]
  rho <- (exp(2 * z) - 1) / (1 + exp(2 * z))
  n <- length(data$y)
  y <- rep(0,n)
  y[1] <- rnorm(1, mu, s = sigma * sqrt(1 / (1 - rho^2)))
  for(i in 2:n)
    y[i] <- mu + rho * (y[i-1] - mu) + rnorm(1) * sigma
  out$y <- y
  return(out)
}

psifcn.mu <- function(theta) theta[1]

psifcn.rho <- function(theta)
{
   z <- theta[3]
   rho <- (exp(2 * z) - 1) / (1 + exp(2 * z))
   return(rho)
}
```

Figure 3. *Functions provided by user for AR(1) example.*

Our package with the functions in Figure 4 provides, for testing that the slope is unity, results $r_\psi = 2.19$ (P=0.014) and $r^*_\psi = 1.98$ (P=0.024). We note that although the total adjustment $r^*_\psi - r_\psi$ is only about -0.2, the NP and INF adjustments are -0.33 and 0.11, with opposite sign.

It can be seen from this that the proposal of Finney results in a standard error that is much too large, under our model. The reason for this was indicated previously in terms of the implicit variance functions for the overdispersion.

```
loglik.binOD <- function(theta, data)
{
  p.bound <- function(p, eps=2.22e-15) p <- (1 - eps) * (p - 0.5) + 0.5
  y <- data$y
  den <- data$den
  X <- data$X
  gq <- data$gq
  n <- length(y)
  p <- ncol(X)
  beta <- theta[1:p]
  sigma <-  exp(theta[p+1])
  linpred <- X %*% beta
  L <- rep(0,n)
  for (i in 1:n)
  {
    prob <- p.bound(plogis(linpred[i] + gq$nodes * sqrt(2) * sigma))
    likq <- y[i] * log(prob) + (den[i] - y[i]) * log(1-prob)
    L[i] <- sum(gq$weights * exp(likq) ) / sqrt(2 * pi)
   }
   return(sum(log(L)))
}

gendat.binOD <- function(theta, data)
{
   out <- data
   den <- data$den
   X <- data$X
   p <- ncol(X)
   n <- length(data$y)
   beta <- theta[1:p]
   sigma <-  exp(theta[p+1])
   u <- rnorm(n) * sigma
   linpred <- X %*% beta + u
   out$y <- rbinom(n, size=den, prob=plogis(linpred))
   return(out)
}
```

Figure 4. *Functions provided by user for binomial overdispersion.*

For evaluating the approximation of our Eq. (1), we simulated 50 000 trials under the hypothesis fit parameters, finding that 2.86% of the r_ψ-values were greater than the observed value of $r_\psi = 2.19$, compared with the 2.4% predicted by $1 - \Phi(r_\psi^*)$ as in (1).

5 Computation of the Jacobian Terms

The challenge in computing r^* involves the sample space derivatives given for (12) and elsewhere in this paper, which are largely Jacobians. The primary difficulty is that in these partial derivatives, some suitable ancillary must be held fixed. This is so difficult that use of this theory was largely stifled for most of the decade 1986–1996. In a major advance, Skovgaard (1996) & Skovgaard (2001) developed a way of approximating these to second order that involves only computing some log likelihood-based covariances computed without conditioning on an ancillary. Although the Skovgaard approach does aim for conditioning on an ancillary, it is compatible with any ancillary meeting the needs for the likelihood ratio approximation of Section 3.

Note that the ancillary information appears in the results through the term $\hat{\imath}^{-1}\,\hat{\jmath}$, which is part of the reason we like to think in terms of the Efron–Hinkley ancillary.

Our aim is to approximate the sample space derivative $\partial^2\ell(\psi,\hat{\lambda}_\psi;\hat{\theta},a)/\partial\theta\partial\hat{\theta}^\top$ and $\partial\{\ell(\hat{\psi},\hat{\lambda};\hat{\theta},a)-\ell(\psi,\hat{\lambda}_\psi;\hat{\theta},a)\}/\partial\hat{\theta}$, that arise in the Jacobians of Section 4. Note that NP_ψ and INF_ψ of (13) can be calculated from those quantities. The Skovgaard approximation to those quantities is given by

$$\begin{aligned}\partial^2\ell(\psi,\hat{\lambda}_\psi;\hat{\theta},a)/\partial\theta\partial\hat{\theta}^\top &\doteq \mathrm{cov}_{\theta_0}\{U(\theta_0),U(\theta)\}\big|_{(\theta_0=\hat{\theta},\,\theta=\tilde{\theta})}\;\hat{\imath}^{-1}\hat{\jmath}\,,\\ \partial\{\ell(\hat{\psi},\hat{\lambda};\hat{\theta},a)-\ell(\psi,\hat{\lambda}_\psi;\hat{\theta},a)\}/\partial\hat{\theta} &\doteq \mathrm{cov}_{\theta_0}\{U(\theta_0),\ell(\theta_0)-\ell(\theta)\}\big|_{(\theta_0=\hat{\theta},\,\theta=\tilde{\theta})}\;\hat{\imath}^{-1}\hat{\jmath}\,,\end{aligned} \tag{14}$$

where the covariances are computed without conditioning on an ancillary. The functions $U(\cdot)$ are ordinary score statistics $U(\theta)=\partial\ell(\theta;y)/\partial\theta$. The final terms $\hat{\imath}^{-1}\hat{\jmath}$ serve to adjust these to conform to ancillary conditioning. This is the Skovgaard (1996) approximation. The error in these approximations is of second order for normal deviations for both the parameter and the ancillary.

It is often unreasonable to compute the required log likelihood-based covariances in (14) exactly, and it is best to approximate them by a simple simulation of datasets under the model. This simulation involves no model fitting and is very different in this respect from a parametric bootstrap. The required number of simulation trials is not greater than a few thousand, because the aim is only estimation of covariances, rather than tail probabilities directly.

For full-rank exponential families, no ancillary must be held fixed for sample space derivatives, and there are closed-form expressions for these sample space derivatives, for example Barndorff-Nielsen and Cox (1994, example 6.24). However, these are obtained exactly with the simulation approach, so it is better for computations not to distinguish between the full-rank exponential family and the general settings. Skovgaard (1996) noted that his approximation is exact for full-rank exponential families, and the type of argument employed by Severini (1999,section 3) shows that the simulation yields the exact Skovgaard covariances. Skovgaard's argument is given in terms of curved exponential families, with his general result being based on approximating other models by that form. For the general setting, the validity of the Skovgaard approximation may be clarified from the more explicit arguments of Severini (1999); see also the two introductions to chapters 9 and 11 in Reid & Martinussen (2017).

6 Topics on Discrete Data and Similar Tests

6.1 Discrete Data and the r^* Approximation

The r^* approximation applies well to discrete data. When r is highly discrete, consideration of continuity correction is of interest, but for the following reason, it can be reasonable not to make such correction.

For approximating literally $P\{r_\psi(Y)\le r_\psi(y);\theta:\psi(\theta)=\psi\}$, a continuity correction should be applied. But for discrete settings, there are reasons to utilise the mid-P, which is the average of the expression just given and the one employing strict inequality. Because of the discreteness, this mid-P has more nearly a uniform distribution under the hypothesis, but there are other reasons to prefer it. It is easily seen that the result of employing r^* without any continuity correction provides an approximation to the mid-P; see Pierce & Peters (1999) and Davison *et al.* (2006).

Considering this more precisely is in principle not complicated. In discrete settings, the relation (1) amounts to approximating a discrete distribution by a continuous one, which involves

standard elementary issues. The most accurate approximations involve continuity correction, which should ideally in principle be performed in terms of the distribution of r or of P-values, rather than the original data. Though there are some general formulae for this, summarised by Pierce & Peters (1992), it is typically simpler and adequate to make the continuity correction to the data before computing r^*. When the inference is conditional on sufficient statistics for the nuisance parameter, the continuity correction should conform to this. These and other matters are clarified in Example 5 that follows.

So, ignoring the continuity correction when employing r^* leads approximately to the mid-P. When the distribution of r is highly discrete, either this or the exact mid-P differs considerably from the exact P-value for discrete data. There are reasons supporting the view that approximating the distribution of r as continuous, without continuity correction, may be preferable to most other methods. In particular, this may be true when the discreteness arises largely from conditioning on nuisance parameter estimates to achieve similar tests. The conditional sample space may be too limited to be useful, or even degenerate. There are many difficulties with similar, or 'exact', tests, some of which are raised in section 5.2 of Cox & Hinkley (1974) and others in Pierce & Peters (1999). It is attractive but rather intractable to sacrifice the exact conditioning in favour of some kind of approximate conditioning, which may involve modest loss of the similarity. Pierce & Peters (1999) argued that in this case, treating the distribution of r as continuous provides a means of approximate conditioning.

6.2 Use of r^* for Approximating the Conditioning for Similar Tests

It is related to the discussion previously given that r^* can be employed to approximate the conditioning on nuisance parameter estimates in full-rank exponential families where ancillary conditioning is not required; this was developed by Pierce & Peters (1992). Jensen (1986) noted that in full exponential families when $\psi(\theta)$ is linear in the canonical parameters, the likelihood ratio statistic $r_\psi(Y)$ is to third order independent of a nuisance parameter estimator. Thus, when the inference is based on this likelihood ratio statistic, the conditioning to achieve similar tests is superfluous. The use of r^* can still be used to obtain a second-order approximation to P-values as in (1). In the package `likelihoodAsy`, the routines do not need to be informed that this is different from ancillary conditioning. The following example illustrates this.

```
loglik.Pois <- function(theta, data)
{
  y <- data$y
  y <- y + 0.50 * c(-1,1,1,-1)
  mu <- exp(data$X %*% theta)
  el <- sum(y * log(mu) - mu)
  return(el)
}

gendat.Pois <- function(theta, data)
{
  out <- data
  mu <-  exp(data$X %*% theta)
  out$y <- rpois(n=4, lam=mu)
  return(out)
}
```

Figure 5. *Functions for* 2×2 *contingency table.*

Example 5. 2×2 contingency table.

Consider testing independence in the 2×2 contingency table with entries $y_{ij} = \{15, 9, 7, 13\}$ where the first and last numbers are the diagonal elements, by conditioning on the marginal totals as usual. This is an instance of the conditioning to eliminate nuisance parameters raised previously. That is, the probability model for the usual 'exact' test arises from conditioning on the row and column totals in a Poisson model. For testing independence, the interest parameter ψ can be taken as the interaction term of the `theta` vector; see the code in Figure 5. The conditioning on the sufficient statistics for the remaining coordinates is performed automatically in the r^* theory presented here, because as noted previously, the marginal distribution in Eq. (11) in this full exponential family setting agrees to third order with the conditional distribution. Because no ancillary conditioning is required, and the example is otherwise simple, it is easy to calculate the exact P-value conditionally on the table margins.

For continuity correction, it is desired to maintain the same row and column totals. The nearest datasets in this sense would have ± 1 added to the diagonal cells and subtracted from the off-diagonal cells. For continuity correction, one might move half-way to that nearest table, adding and subtracting 0.50 to cells. The exact one-sided P-value is 0.0646, and the r^*_ψ P-value, for testing that the odds ratio is unity, with such continuity correction on the original data is 0.0676. This agreement is the main point we are making, but to continue, the mid-P is 0.0404 and the r^*_ψ P-value without continuity correction is 0.0362. This agreement is less precise, but the argument for it given by Pierce & Peters (1999) is more approximate than for the other comparison. Our preference is the P-value without continuity correction, but the other numbers are of some interest.

It is not difficult to generalise this example to the case of several nuisance parameters, which arise for example in a collection of $k \times 2$ tables with common interaction parameter in a logistic regression model, as would arise in case–control studies with $k : 1$ matching. See for example Brazzale *et al.* (2007), sections 9.2 and 9.3, reporting results for D.R. Cox's famous *Crying Babies* data, where a stratified logistic regression model can be applied. This is an instance where the methods of this paper can be adequate even for a large number of strata, especially so when the stratum sizes are not minimally small.

7 Discussion and Conclusion

The theory and methods presented here reflect an important chapter in the developments in statistics. The approximations can be considerably better than usual first-order ones. The development is also an important complement to the Neyman–Pearson theory, which many, particularly outside of the USA, find limited by the emphasis on decision-making; see for example Reid (2005, section 3). Regardless of one's view on this, approximately optimal inference has not been fully integrated into the Neyman–Pearson theory.

The theory of ancillary conditioning is no doubt demanding for a general statistical audience, but we have attempted here to make this more accessible than in the original developments. As with other theoretical advances, success hinges on the existence of suitable software. So this paper has the dual aims of useful exposition and announcement of an `R` package applying the methods for a quite general setting.

It will have occurred to some readers that the terminology 'Modern Likelihood-Frequentist Inference' presumes there will be considerable further accessibility and acceptance of the methodology dealt with in the exposition here. We hope this paper achieves the aim of facilitating this.

Endnotes

Endnote 1

Unless considerable restraint is employed, the theory of ancillarity is quite complicated, with grounds for pessimism in its practical value. A useful discussion of these difficulties is given by Ghosh *et al.* (2010). However, as they acknowledge, many of the difficulties can become less severe in terms of approximate ancillarity. The main issues for present purposes are as follows: (i) Skovgaard (1985) and others have shown that under quite general regularity conditions, the ancillary $a = \Gamma(\hat{\theta})\,\hat{\imath}^{-1}\hat{\jmath}$ of Section 2 meets the requirements (iii) and (iv) of Section 1.2, although not uniquely and (ii) this ancillary conforms well to the requirements for the likelihood ratio approximation of Section 3. Efron & Hinkley (1978) had earlier shown more discursively the value of this ancillary, conjecturing some of the results established by Skovgaard, who obtained more rigorous results on this.

Endnote 2

The issues of second-order and third-order, that is $O(n^{-1})$ and $O(n^{-3/2})$, results are subtle aspects of modern likelihood asymptotics. One distinction in this is that often the distribution of r^*_ψ is standard normal to third order, whereas more inferentially, clear results on the distribution of r_ψ as in (1*) are generally valid only to second order. This is indicated in eq. (7.4) of Severini (2000). The distinction also arises in comparing the approximations to sample space derivatives due to Fraser and colleagues; for example Fraser (2004), Reid & Fraser (2010) and that due to Skovgaard (1996) employed for this paper. For approximating the distribution of r_ψ, as we emphasise in this paper, both approaches are generally of second order for reasons just stated. However, in general, the 'large deviation' property discussed in relation to Eq. (1*) is far more important than whether a result is of second or third order. Both the Fraser and Skovgaard approaches have the large deviation property. There are other second-order methods that do not, most often those based on 'orthogonal parameters'; Cox & Reid (1987), Severini (2000, section 7.5.2).

Endnote 3

The likelihood ratio approximation to the density of $\hat{\theta}$ is more often referred to as the *saddlepoint approximation* or *Barndorff-Nielsen's p^* formula*. The term saddlepoint refers to inversion of the cumulant generating function, involving a saddlepoint in the complex coordinate system: for example Daniels (1954). Reid (1988) provides a review of extensive work following that development, involving far more workers than we can mention here. For full-rank exponential families, the approximate distribution of the maximum likelihood estimator was obtained through a more statistical approach without complex analysis, culminating in the work of Barndorff-Nielsen & Cox (1979). For location-scale models, where the estimator is generally not sufficient, Fisher (1934) had obtained a result of this nature on its distribution by conditioning on the configuration ancillary.

Endnote 4

In multiparameter problems, definition and calculation of the terms involved in r^*_ψ is rather intricate, even though their basis is readily grasped. These intricacies are presented succinctly in eqs 6.102–6.108 of Barndorff-Nielsen & Cox (1994). Even experienced analysts often find it difficult to get these correctly, when doing each problem from a fresh start. A considerable part of our motivation in providing the `R` package is to remove the need for users to master these potential difficulties.

Acknowledgments

This research was initiated while Donald Pierce was visiting the University of Padova in 2012, funded by a Visiting Scientist Fellowship awarded to Alessandra Salvan by the University of

Padova. This work was also supported by a grant from the Italian *Ministero dell'Istruzione, dell'Università e della Ricerca*. The authors are grateful to Dawn Peters for suggestions that improved the exposition, to Luigi Pace and Ib Skovgaard for insightful comments and to Thomas Severini for answering our technical questions without fail. We are also grateful to an anonymous referee for valuable suggestions.

References

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**(2), 343–365.

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, **73**(2), 307–322.

Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika*, **78**(3), 557–563.

Barndorff-Nielsen, O. E. & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *J. Roy. Statist. Soc. Ser. B*, **41**(3), 279–312.

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

Brazzale, A. R. & Davison, A. C. (2008). Accurate parametric inference for small samples. *Statist. Sci.*, **23**(4), 465–484.

Brazzale, A. R., Davison, A. C. & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics.* Cambridge: Cambridge University Press.

Butler, R. W. (2007). *Saddlepoint Approximations with Applications* Cambridge: Cambridge Univ. Press.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics* London: Chapman and Hall.

Cox, D. R. (1980). Local ancillarity. *Biometrika*, **67**(2), 279–286.

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

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.*, **25**(4), 631–650.

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *J. Roy. Statist. Soc. Ser. B*, **50**(3), 445–461.

Davison, A. C. (2003). *Statistical Models.* Cambridge: Cambridge Univ. Press.

Davison, A. C., Fraser, D. A. S. & Reid, N. (2006). Improved likelihood inference for discrete data. *J. Roy. Statist. Soc. Ser. B*, **68**(3), 495–508.

Davison, A. C., Hinkley, D. V. & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statist. Sci.*, **18**(2), 141–157.

DiCiccio, T. J., Martin, M. A. & Stern, S. E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canad. J. Statist.*, **29**(1), 67–76.

DiCiccio, T. J. & Young, G. A. (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika*, **95**(3), 747–758.

DiCiccio, T. J., Kuffner, T. A., Young, G. A. & Zaretzki, R. (2015). Stability and uniqueness of p-values for likelihood-based inference. *Statist. Sinica*, **25**, 1355–1376.

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika*, **67**(2), 311–333.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika*, **65**(3), 457–487.

Efron, B. (1998). R. A. Fisher in the 21st century (with discussion). *Statist. Sci.*, **13**(2), 95–122.

Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**(4), 826–838.

Finney, D. J. (1947). *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve.* Cambridge: Cambridge Univ. Press.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. A*, **144**(852), 285–307.

Fraser, D. A. S. (1991). Statistical inference: likelihood to significance. *J. Amer. Statist. Assoc.*, **86**(414), 258–265.

Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.*, **19**(2), 333–369.

Fraser, D. A. S. & Reid, N. (1988). On conditional inference for a real parameter: a differential approach on the sample space. *Biometrika*, **75**(2), 251–264.

Fraser, D. A. S., Reid, N. & Sartori, N. (2016). Accurate directional inference for vector parameters. *Biometrika*, **103**(3), 625–639.

Ghosh, M., Reid, N. & Fraser, D. A. S. (2010). Ancillary statistics: a review. *Statist. Sinica*, **20**, 1309–1332.

Jensen, J. L. (1986). Similar tests and the standardized log likelihood ratio statistic. *Biometrika*, **73**(3), 567–572.

Johansen, S. (1995). The role of ancillarity in inference for non-stationary variables. *Econ. J.*, **105**(429), 302–320.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *Statistical Analysis of Failure Time Data*, 2nd ed. Hoboken, NJ: Wiley.
Lawless, J. F. (1973). Conditional versus unconditional confidence intervals for parameters of the Weibull distribution. *J. Amer. Statist. Assoc.*, **68**(343), 665–669.
Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. Hoboken, NJ: Wiley.
Lozada-Can, C. & Davison, A. C. (2010). Three examples of accurate likelihood inference. *The Amer. Statist.*, **64**(2), 131–139.
McCullagh, P. (1984). Local sufficiency. *Biometrika*, **71**(2), 233–244.
McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008). *Generalized Linear and Mixed Models*, 2nd ed. New York: Wiley.
Nocedal, J. & Wright, S. (2006). *Numerical Optimization*, 2nd ed. New York: Springer-Verlag.
Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective* Singapore: World Scientific Publishing.
Pierce, D. A. & Bellio, R. (2006). Effects of the reference set on frequentist inferences. *Biometrika*, **93**(2), 425–438.
Pierce, D. A. & Bellio, R. (2015). Beyond first-order asymptotics for Cox regression. *Bernoulli*, **21**(1), 401–419.
Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc. Ser. B*, **54**(3), 701–737.
Pierce, D. A. & Peters, D. (1999). Improving on exact tests by approximate conditioning. *Biometrika*, **86**(2), 265–277.
R Core Team. (2017). *R: A Language and Environment for Statistical Computing* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Reid, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statist. Sci.*, **3**(2), 213–238.
Reid, N. (1996). Likelihood and higher-order approximations to tail areas: a review and annotated bibliography. *Canad. J. Statist.*, **24**(2), 141–166.
Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, **31**(6), 1695–1731.
Reid, N. (2005). Asymptotics and the theory of statistics. In *Celebrating Statistics: Papers in Honour of D.R. Cox*, Eds. A. C. Davison, Y. Dodge & N. Wermuth, pp. 73–88. Oxford: Oxford University Press.
Reid, N. & Fraser, D. A. S. (2010). Mean loglikelihood and higher-order approximations. *Biometrika*, **97**(1), 159–170.
Reid, N. & Martinussen, T. (2017). *Inference, Asymptotics, and Applications: Selected Papers of Ib Michael Skovgaard, with Introductions by his Colleagues,* Eds. N. Reid & T. Martinussen, New Jersey: World Scientific Publishing.
Severini, T. A. (1990). Conditional properties of likelihood-based significance tests. *Biometrika*, **77**(2), 343–352.
Severini, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika*, **86**(2), 235–247.
Severini, T. A. (2000). *Likelihood Methods in Statistics.* Oxford: Oxford University Press.
Skovgaard, I. M. (1985). A second-order investigation of asymptotic ancillarity. *Ann. Statist.*, **13**(2), 534–551.
Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, **2**(2), 145–165.
Skovgaard, I. M. (2001). Likelihood asymptotics. *Scand. J. Statist.*, **28**(1), 3–32.
Sweeting, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika*, **82**(1), 1–23.
Young, G. A. (2009). Routes to higher-order accuracy in parametric inference. *Aust. N.Z. J. Stat.*, **51**(2), 115–126.
Young, G. A. & Smith, R. L. (2005). *Essentials of Statistical Inference.* Cambridge: Cambridge University Press.

[*Received April 2016, accepted June 2017*]