Modern Likelihood-Frequentist Inference

Donald A. Pierce

Oregon Health and Sciences University

Portland, Oregon U.S.A

Ruggero Bellio

University of Udine

Udine, Italy

Summary

We offer an exposition of modern higher-order likelihood inference, and introduce software to implement this in a fairly 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 the extra-likelihood aspects of 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, and transformation of this yielding a second-order approximation to the distribution of the likelihood ratio test statistic. This follows developments of Barndorff-Nielsen and others. The software utilizes the approximation to required Jacobians as developed by Skovgaard, which is included in the exposition. Several examples of using the software are provided.

Some Key Words: ancillary statistic, conditional inference, likelihood asymptotics, modified likelihood ratio, modified profile likelihood, neo-Fisherian inference, nuisance parameter, saddlepoint approximation, p^* formula.

1. INTRODUCTION AND BASIC CONCEPTS

Special likelihood-based procedures, modifying usual inferential approximations for much higher accuracy, have emerged in recent years (Davison, 2003, Ch. 12; Brazzale and Davison, 2008; Lozada-Can and Davison (2010). The performance of these methods is uncanny, and it is not unreasonable to characterize them as often 'close to exact'. 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 is includes both an exposition aiming to make more accessible the main ideas of this development, and provides novel software tools for implementing them.

These tools are embodied in the package likelihoodAsy in R (**R** Core Team, 2016), which is available at the Comprehensive R Archive Network (CRAN) *http://cran.r-project.org/* . This package applies very generally, requiring primarily a user-supplied function for evaluating the likelihood. However, inferences beyond first order require model specification beyond the likelihood function, which is achieved by another user-provided function that generates a sample under the model.

We have particular motivations for providing the exposition 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 wide grasp of these developments, as might find its way into textbooks. This endeavor was initiated by Young and Smith (2005*).* Without this wider grasp, 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, e.g. Brazzale, Davison and Reid (2007), Brazzale and Davison (2008), Lozada-Can and Davison (2010). Accompanying the text by Brazzale, Davison and Reid were software tools of a different nature than here. We do not attempt to document fully the theory developed in this paper, involving too many contributions to be done briefly, but as noted very useful reviews were provided by Reid.

There are largely two limiting issues regarding the adequacy of first-order methods: (a) limited information on the interest parameter, and (b) effects of fitting nuisance parameters. Issue (a) is the 'sample size' matter that would first come to mind in considering this, but (b) can be important even for moderately large samples. Thus it can be said that (b) is often more practically important than (a), though the latter is certainly theoretically important. The software here provides diagnostics that assess these two matters, which would be less clear from direct simulation.

We will say that 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 and Reid (1988), Fraser (1991, 2004), Reid (2003), Brazzale and Davison (2008). The closest point of contact with the partic-

ulars of this paper arises in the approximations considered in Section 5, where we introduce an Appendix regarding the Fraser approach. Anthony Davison, starting with Davison (1988) has done much to promote the ideas reviewed in this paper and particularly for approximate conditional inference, including what we introduce in EndNote 1; see Chapter 12 of Davison (2003) and the three citations of the previous paragraph with Davison as co-author.

Focus in these methods is on testing a hypothesis on the value of a scalar function $\psi(\theta)$ of the model parameters, with confidence intervals obtained by numerically inverting such tests. Let $W_\psi(y)$ be the usual generalized likelihood ratio statistic, detailed below, with limiting χ^2_1 distribution, and consider directional inference based on $r_\psi(y) = sign(\hat{\psi} - \psi)\sqrt{W_\psi(y)}$. Then for observed data y, a one-sided version of the usual first-order inference can be based on the result

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

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

$$p\{r_\psi(Y) \le r_\psi(y) : \psi(\theta) = \psi\} = \Phi\{r^*_\psi(\mathrm{y})\}\{1 + O(n^{-1})\}\ , \qquad (1)$$

along with further properties discussed in the next section. 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 $|i|$. 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.

The reason for focus on the signed root of the likelihood ratio is so that the inference can capture the direction of departure from the hypotheses. Optimal two-sided confidence intervals must involve complex decisions regarding the allocation of error rate between the tails. As suggested by Cox & Hinkley (1974, Sect 4.7) it is preferable to take two-sided confidence intervals as the combination of one-sided regions with a specified allocation between these two error rates.

The quantity r^* was derived by Barndorff-Nielsen (1986, 1991) in path-breaking work, but in a form difficult to compute in general. Various approximations have emerged, and in this paper and the accompanying software we utilize the version developed by Skovgaard (1996, 2001). The work of Fraser and colleagues referred to above led to a different version of r^*, as noted above. Sometimes workers distinguish notationally between the original r^* and approximations to it; e.g. 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), Sect 7.5), but are inferior to that employed here for reasons explained later. We utilize simple simulation, without model fitting, to apply the Skovgaard method, but do not consider that as yet another approximation, but only a way to facilitate broad application of Skovgaard's remarkable advance.

Understanding the basis for (1) emphasizes 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-maximizing 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:

(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, e.g. 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-maximization of the Neyman-Pearson theory. This emphasis on sufficiency is a key aspect of Neo-Fisherian inference, 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 , i.e. its distribution is free of θ , and

(iv) $(\hat{\theta}, a)$ is sufficient, and conditionally on a , the estimator $\hat{\theta}$ is sufficient.

Though the development in this paper emphasizes the general setting where $\hat{\theta}$ is not sufficient, which requires the conditioning in (*iii-iv*), the r^* results and our package likelihoodAsy are also useful when the maximum likelihood estimator is sufficient, i.e in full-rank exponential families. In that case the conditioning is not on an ancillary, but rather is for a standard approach for 'eliminating' nuisance parameters, achieving what in the Neyman-Pearson approach are called *similar tests*. See EndNote 1.

This conditioning of *(iii-iv)* is most important in the considerations of this paper. Generally, the ancillarity is almost always approximate; see Endnote 2. The results of conditioning on any such second-order ancillary are to that order unique. An important version of that, which always exists under mild regularity conditions, is the Efron-Hinkley ancillary introduced in Sect. 2, which is essentially the ratio of observed to expected information. The most important consequence of such conditioning is that it can 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-maximizing Neyman-Pearson theory. This encapsulates what has been termed Neo-Fisherian inference; see Pace and Salvan (1997).

Before turning to further details of general issues we offer an example 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 $S(t_i;\beta,\gamma)=\exp[-\{t_i^{\gamma}\exp(z_i\beta)\}]$, 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 for the survival probability or reliability, at a given time t_0 , namely $S(t_0;\beta,\gamma)$ and for a specified covariate vector z_0 . The interest parameter ψ will be represented as the log reliability

$\psi(\beta,\gamma)=-t_0^{\gamma}\exp(z_0\beta)$, where using this logarithmic representation does not affect r_{ψ}^* , but it can affect the numerical behavior of the constrained maximization routine.

For the Weibull model the 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 (Lawless 1973; Lawless 2003 Appendix E, Example 5.21 of Davison (2003), and though exact inference conditional on this ancillary can be accomplished with one-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 leukemia survival. We will choose t_0 and z_0 such that, for our dataset, the maximum likelihood estimate of the reliability is 0.10. For an hypothesis on this with -values small enough to be of interest, we will test that the reliability is 0.03, which is approximately a lower 95% confidence limit based on first-order methods, against the alternative of larger values.

The functions to be provided by the user of our software are shown in Figure 1. We note that although allowing for censored data would entail only a minor change in the likelihood routine, the data-generation routine would need to involve a probability model for the censoring, which is daunting. The most commonly-used models for this are not really intended to be realistic, and there are interesting issues that need further consideration when going beyond first-order asymptotics where this model matters to the inference; these issues have been considered by Pierce and Bellio (2015).

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

$$r_\psi = 1.67 \ \ (P = 0.048)$$
$$r_\psi^* = 2.10 \ \ \left(P = 0.018\right)$$
$$Wald = 1.96 \ (P = 0.025) \ ,$$

where the latter is the Wald statistic in representation $(\hat{\psi} - \psi)\} / SE(\hat{\psi})$. Confidence limits are shown in Figure 2. It was mentioned 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.

```
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

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_ψ, using Weibull datasets with parameter values fitted to the analysis dataset under the hypothesis on ψ. The result with 10,000 simulation trials is that, empirically, $p\{r_\psi(Y) > r_\psi(y) : \psi = \log(0.03)\} = 0.019$. This compares favorably to $\Phi(-r_\psi^*) = 0.018$, whereas $\Phi(-r_\psi) = 0.048$.

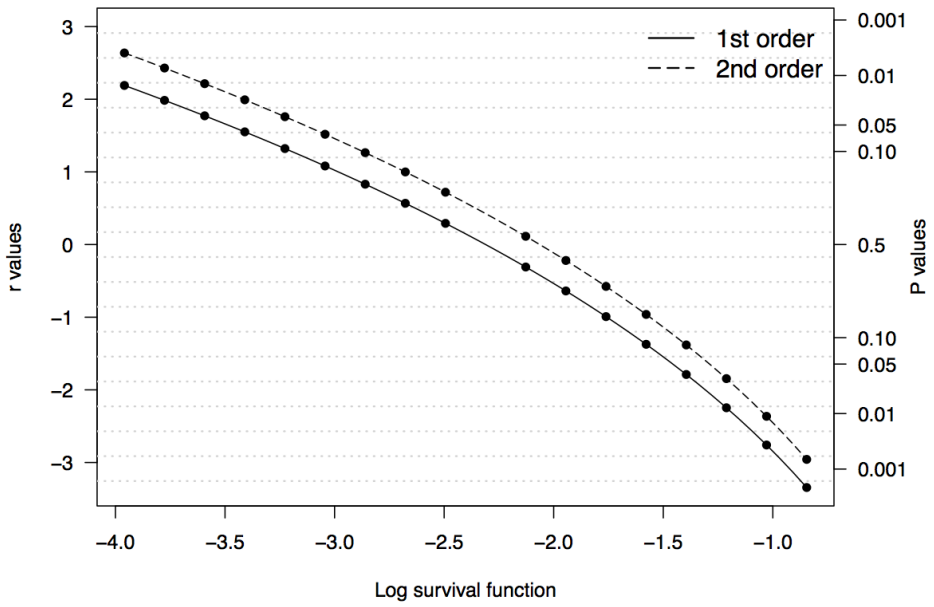

Figure 2. One-sided confidence limits at all levels, using 1st and 2nd order likelihood ratio methods. The *P*-values are the one-sided error rates.

We now briefly consider another example emphasizing 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. Veteran's Administration Cancer Data.

These data are from Appendix A of Kalbfleisch & Prentice (2002), and are also in the dataset veteran of the R package survival. 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 4 covariables: *Karnofski performance*, *diagnosis time*, *age* and *prior therapy*. Inference here regards testing that the survival function at 400 days is > 0.2, for covariate values 60, 7, 69 and 0. For a Weibull model analysis the user-defined functions can be as for Example 1, with obvious minor changes for defining the interest parameter and also allowance for censored data. The *P*-values are: 0.033 based on r and 0.079 based on r^*, even though the sample size is substantial. About 70% of the inadequacy of the first-order result results from fitting the 6 nuisance parameters (including the constant term and shape parameter), with the remainder being due to limited information.

2. GENERAL PRELIMINARY ISSUES

Let $pr(y;\theta)$ be the density (or probability mass function) for a dataset y that is not necessarily a collection of independent observations, with $\dim(\theta) = p \geq 1$. The observations can be continuous or discrete, and the primary regularity condition is that $pr(y;\theta)$ is a differentiable function of the parameter, ruling out typical settings where the set of *y*-values where $pr(y;\theta) > 0$ depends abruptly on θ . The likelihood function $L(\theta; y)$ is any function that is proportional to $pr(y;\theta)$ for given *y,* and we write $l(\theta; y)$ for the loglikelihood function. The *observed information* matrix for observed data *y* is $j(\theta) = -\partial^2 l(\theta; y) / \partial\theta\partial\theta^T$, and we write $\hat{j}$ for the matrix $j(\hat{\theta})$. When $\hat{\theta}$ is not a sufficient statistic, $\hat{j}$ 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 scalar parametric function $\psi = \psi(\theta)$, which is referred to as the *interest parameter*. It is convenient to utilize 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 in Example 1 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 and other functions of the likelihoodAsy package. We will first assume that the distribution of y is continuous, and then deal in Section 6 with discrete settings.

The directional likelihood ratio statistic is defined as

$$r_\psi = sign(\hat{\psi} - \psi)\sqrt{2\{l(\hat{\theta}; y) - l(\tilde{\theta}; y)\}}\ , \qquad (2)$$

where $l(\theta; y)$ is the loglikelihood function, $\hat{\theta}$ is the maximum likelihood estimator, and $\tilde{\theta} = (\psi, \hat{\lambda}_\psi)$ 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 suitable for directional inference, and as described in Section 1 the random variable r_ψ has to $O_P(n^{-1/2})$ 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 second-order ancillary will meet the needs for this paper. The main result in (1) and (1*) below 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 i.e. the matrix $\hat{i}^{-1}\hat{j}$. In order for the distribution of the ancillary to be more nearly constant in θ, we may rescale $\hat{i}^{-1}\hat{j}$ by dividing it, in a matrix sense, by $n^{1/2}$ times its estimated asymptotic standard deviation, resulting in an ancillary statistic we will refer to by a. This ancillary, reflecting ideas of Fisher and others, is called the Efron-Hinkley ancillary, studied in the paper Efron and Hinkley (1978) that is notable for its exposition.

Note that $\dim(a)$ does not depend on the sample size. The scaling employed above renders $a = O_P(n^{-1/2})$, this being an instance of what Barndorff-Nielsen (1984) termed an *intrinsic* representation. Such representations are useful for keeping in mind the important issue that the effect of using ancillary information is of $O_P(n^{-1/2})$, as shown in Barndorff-Nielsen (1984). Intrinsic representations must, of course, be rescaled by $n^{1/2}$ in order to have a limiting distribution.

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.

We can now express (1) somewhat more rigorously as follows. Let Y have distribution given by $pr(y;\psi,\lambda)$, y denote the observed dataset, and a denote the Efron-Hinkley or another locally second-order ancillary. Then for testing an hypothesis on $\psi = \psi(\theta)$,

$$p\{r_\psi(Y) \le r_\psi(y) \mid a;\psi,\lambda\} = \Phi\{r_\psi^*(y)\}\{1+O(n^{-1})\}, \qquad (1^*)$$

for $\|\hat{\psi}-\psi\| = O(n^{-1/2})$. The result (1*) remains valid, as in (1), without conditioning on the ancillary a, as it is a key feature of likelihood ratio statistics that they are to second order independent of any ancillary; McCullagh (1984), Severini (1990). Note the implication that for moderate deviations, the *P*-values referred to by (1*) do not, to second order, depend on the nuisance parameter. The error bound can be expressed more clearly and comprehensively as $\{1+O_P[\|\hat{\psi}-\psi\|\times\|a\|]\}$, where the ancillary is represented intrinsically so that $\|a\| = O_P(n^{-1/2})$. These conditions on the parameter and ancillary are referred to as specifying 'moderate deviations', If one, but not both, of the parameter and ancillary deviations is not moderate, then the error in the expression (1*) becomes $\{1+O(n^{-1/2})\}$, and for many second-order methods that measure of the error is not this small; see EndNote 3. This 'protection' for large deviations is crucial for second-order results to be most useful, since that protection

results in good approximation at the extremes of moderate deviations. Skovgaard (1996, 2001) proved these claims, for the version of r^* we employ here, though with weaker specification of the ancillary as noted above.

Since (1*) pertains to the true distribution of $r_\psi(Y)$, an alternative approach to computing *P*-values would involve direct simulation. An important issue is that a standard unconditional simulation leads to inferences agreeing with the ancillary conditioning in (1*), since the likelihood ratio statistic is to second order independent of any ancillary; see Davison, Hinkley and Young (2003), DiCiccio, Martin and Stern (2001), DiCiccio, Kuffner and Young (2015), Young (2009). This asymptotic independence also obtains with other pivotals agreeing sufficiently well with the likelihood ratio, in particular the Wald statistic using observed information. It is well known and considered in the references just given, that if the constrained maximum likelihood estimator $\hat{\lambda}_\psi$ 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 $r_\psi(Y)$. However, this does not apply to the ancillary conditioning, where the approximation remains $O_P(n^{-1})$, as is the case when using the unconstrained maximum likelihood estimator $\hat{\lambda}$. These approaches are often referred to as the *parametric bootstrap.*

That the parametric bootstrap provides to second order the same *P*-values obtained from r^* does not lessen the interest in using the latter. For one thing, the theory underlying this is simply an important development in statistical theory. For another, this provides useful diagnostics on the shortcomings or first-order inferences that are not revealed by the parametric bootstrap, Finally, the bootstrap involves repeated fitting of the model, which can be problematic. The fitting procedure used by *likelihoodAsy* has been carefully selected to reduce the difficulties in fitting the constrained model under the hypothesis.

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 4. What we call the likelihood ratio approximation is often called Barndorff-Nielsen's p^* formula (Barndorff-Nielsen 1983; Barndorff-Nielsen and Cox 1994). The argument as summarized here is in detail heuristic, and we comment on that afterward.

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

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

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 $p(y \mid \hat{\theta};\theta) / p(y \mid \hat{\theta};\hat{\theta})$ is unity,

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

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 $p(\hat{\theta};\hat{\theta}) = (2\pi)^{-p/2} \mid j(\hat{\theta})|^{1/2}$, 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 normalizing 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 for 'large deviations' of $\hat{\theta} - \theta$. It is not so much that approximations far outside of the 'moderate deviation' region in which $\| \hat{\theta} - \theta \| = O(n^{-1/2})$ are of practical importance, but that the large deviations protection leads to approximations are good even in the extremes of that moderate region. This is often not the case for second-order approximations that are based on expansions around a fixed parameter value; see EndNote 3.

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 $\hat{\theta} \mid a$. 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 ancil-

lary 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 $\hat{\theta} \mid a$. 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), Skovgaard (1985). In a slightly different statement, Reid (1988, Section 3) notes that our results in this section hold for any second-order ancillary.

Due to these approximate sufficiency considerations, the same argument as above applies conditionally, leading now to the same approximation formula, but *interpreted as* approximating the density conditional on an ancillary,

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

Note that in the argument (5) the omitted term $p(y \mid \hat{\theta}, a;\theta) / p(y \mid \hat{\theta}, a;\hat{\theta})$ is no longer exactly unity, but is $1+O(n^{-1})$ due to 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' $j(\hat{\theta})$ 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"*, or simply as "*the magic formula*", e.g. Efron (1998).

It is known that (5), after rescaling to integrate to unity, is exact for location-scale and all other transformation models. This was realized 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 loglikelihood at $\hat{\theta}$ would be asymptotically sufficient.

Reasons to consider the above 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 the desired result in its full generality; see for example Reid (2003), Section 2.2.

The approximations (3) and (5) are of limited practical value for direct use, since 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 these approximations become 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,

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

This is not convenient to use, and we make a further second-order approximations as follows. Note that $\partial r_\theta / d\hat{\theta} = [\partial\{l(\hat{\theta};\hat{\theta},a) - l(\theta;\hat{\theta},a\} / \partial\hat{\theta}] / r_\theta$, from differentiating r_θ^2 in definition (2). This Jacobian term indicates why *sample space derivatives*, i.e. 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 the sample-space derivative $\partial\{l(\hat{\theta};\hat{\theta},a) - l(\theta;\hat{\theta},a\} / \partial\hat{\theta}$ by u_θ, so (6) can be expressed as

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

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 we have $r_\theta^* = r_\theta + O_P(n^{-1/2})$, and we note that $u_\theta / r_\theta = 1 + O_P(n^{-1/2})$ for $|\hat\theta - \theta| = O(n^{-1/2})$. Sweeting (1995) analyzed 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 second-order term $(u_\theta / r_\theta)^2$ we have

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

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),

$$p(r_\theta(Y) \le r_\theta(y) \mid a;\theta) = \Phi[r_\theta^*(y)]\,[1 + O(n^{-1})] \quad . \tag{10}$$

It is somewhat more common to consider the higher-order distribution of r^* , as noted in Sect. 1 and discussed in EndNote 3.

The result (10) as stated holds under the moderate deviation condition $|\hat\theta - \theta| = O(n^{-1/2})$, which is required for the likelihood ratio approximation to the distribution of $\hat\theta$ and other steps following (7). See EndNote 3 regarding second 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. Recall that the likelihood ratio approximation to the density of $\hat\theta \mid a$ applies in the case that $\dim(\theta) > 1$. Thus the changes from the argument above are that in transforming from $pr^*(\hat\theta \mid a;\theta)$ 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 . It is common to think of dealing with nuisance parameters through a further conditioning, at least in exponential families, rather than such marginalization, which was discussed in EndNote 1 and to which we return in Section 6 .

The standard approach for results in this section is to express the approximate marginal distribution of r_ψ in the form

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

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 $\partial^2 l(\hat{\theta}_\psi)/\partial\lambda\,\partial\hat{\lambda}^T$. The sample-space derivative raised for (7) now becomes $\partial\{l_P(\hat{\theta};\hat{\theta},a) - l_P(\theta;\hat{\theta},a\}/\partial\hat{\psi}$, where $l_P(\cdot)$ denotes the profile loglikelihood. 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 $p^*(r_\psi \mid a;\theta)$ 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 and Cox (1994), Sections 6.6 and 6.6.1. The result is again our formula (8), but with a more general definition of r_ψ^*, that can be expressed as

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

where

$$C_\psi = \left|\frac{\partial^2 l(\hat{\theta}_\psi)}{\partial\lambda\,\partial\hat{\lambda}^T}\right|\left\{|\,\hat{j}_{\lambda\lambda}\,||\,\tilde{j}_{\lambda\lambda}\,|\right\}^{-1/2}\,, \quad \tilde{u}_\psi = |\,\partial\{l_P(\hat{\theta};\hat{\theta},a) - l_P(\theta;\hat{\theta},a)\}/\partial\hat{\psi}\,||\,\tilde{j}_{\psi|\lambda}\,|^{-1/2}\,.$$

Here the tilde denotes evaluation at $\{\psi, \hat{\lambda}_\psi\}$, $\tilde{u}_\psi$ is given the sign of r_ψ, and $\tilde{j}_{\psi|\lambda}$ denotes the adjusted information for ψ. The two final terms in (12) derive almost entirely from the Jacobian $|\partial\hat{\theta}/\partial(r_\psi, \hat{\lambda}_\psi)|$, 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 penultimate term is a new object corresponding to Barndorff-Nielsen's *modified profile likelihood* $L_{MP}(\psi;y) \propto L_P(\psi;y)\exp(C_\psi^{-1})$ for the setting of a scalar parameter ψ. This modified likelihood pertains specifically to allowing for effects of fitting nuisance parameters λ. The more general setting of vector interest parameter is considered further in the final section.

It is useful and common to express (12) as

$$r_\psi^* = r_\psi + NP_\psi + INF_\psi\,, \tag{13}$$

referring to the terms as the *nuisance parameter* and *information* adjustments (Pierce and Peters 1992,

Barndorff-Nielsen and 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; i.e. the adjusted information $j_{\psi|\lambda} = j_{\psi\psi} - j_{\psi\lambda} j_{\lambda\lambda}^{-1} j_{\lambda\psi}$ 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 modified profile likelihood introduced just above can be expressed exactly as $L_{MP}(\psi; y) \propto L_P(\psi; y) \exp(-r_\psi \, NP_\psi)$, where L_P denotes the profile likelihood. This is only for the case that $\dim(\psi) = 1$, and the case where $\dim(\psi) > 1$, where *NP* and *INF* are not defined, is considered in the final section. Note that for $\dim(\psi) = 1$ the modified profile likelihood function does not involve the *INF* adjustment, with result that it is not valid to second order in approximating the ideal conditional likelihood when that is available. 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 modified profile likelihood 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 the pseudo-likelihood $\exp\{-(r^*)^2/2\}$.

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

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

with independent Gaussian errors satisfying $\text{var}(\varepsilon_i) = \sigma^2$. 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 done in psifun.rho rather than in likAR1, since the optimization is done on the *theta* scale. Even though our approach is in principle unaffected by choice of parametrization, it makes the optimization 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 constrains by a log transform.

Higher-order likelihood inference for this example was considered by Lozada-Can and Davison (2010). An interesting aspect of this is that in order to apply the Fraser approach of the Appendix, they needed to utilize a special ancillary based on a martingale representation of the AR(1) process, in contrast to the usual one for independent observations, or the general ancillary based on $\hat{i}^{-1}\hat{j}$ that we have in mind here. For their dataset of 48 observations of luteinizing hormone levels measured at 10 minute 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 3 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 $n = 48$ 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 the correlation ρ of 0.765 based on r is found to be an 88% limit based on r^*. The maximum likelihood estimator of ρ is 0.574 ± 0.116. For testing $\rho = 0.765$ we find $r = -1.643$ $(P = 0.050)$ and $r^* = -1.155$ $(P = 0.124)$. The *NP* and *INF* adjustments are 0.36 and 0.13. This r^*-based *P*-value was confirmed by simulation of 10,000 trials using the parametric bootstrap method discussed at the end of Sect. 2.

At the suggestion of a referee we considered reproducible simulated data with the true ρ of 0.95, much closer to the unity value where the process becomes non-stationary. The non-stationary theory for inference strongly involves ancillary conditioning on the observed information, which is based on the sample variance of the *y*-values on which one regresses to fit the model; see Johansen (1995).

With these data the maximum likelihood estimate of ρ is 0.928 and the Wald-based first order 95% upper confidence limit is 0.9998. For testing that value, the distinction between first-order and second-order inferences is, as expected, much greater. The Wald *P*-value is 0.05 as planned, the P-value based on *r* is 0.004, and that based on r^* is 0.1612. The values of *NP* and *INF* become 1.27 and 0.36. We confirmed the *P*-value result by employing the parametric bootstrap simulation of Sect. 2, obtaining a *P*-value of 0.161.

```
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

EXAMPLE 4: Binomial overdispersion

We now consider one of the standard models for overdispersion in binomial data; namely that $\log\{p_i/(1-p_i)\} = z_i\beta + u_i$, where the u_i are independent $N(0,\sigma^2)$ random variables. This form of modeling, now widely used in more general settings as seen in McCulloch, Searle and Neuhaus (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 loglikelihood is to be used, it is particularly important that this numerical integration be highly accurate. This led us to using the Gauss-Hermite quadrature of the code given here, as indicated by the *gq* objects that are included in the dataset, as indicated in the vignette of the HTML 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 since the residual deviance is 36.25 on 8 d.f. , presenting evidence for rather 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 above.

Our package with the functions in Figure 4 provides, for testing that the slope is unity, results $r = 2.19$, $P = 0.014$ and $r^* = 1.98$, $P = 0.024$. We note that although the total adjustment $r^* - r$ is only about 0.2 , the *NP* and *INF* adjustments are -0.34 and 0.55, each larger in magnitude than the net adjustment.

```
loglik.binOD <- function(theta, data)
{
  p.range<- function(p, eps=2.22e-15)
   {
      out <- p
      out[p<eps] <- eps
      out[p>(1-eps)] <- (1-eps)
      return(out)
   }
  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.range(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(log(prod(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

A matter of interest is how these results compare to Finney's suggestion. In terms of Wald statistics for testing slope unity, Finney's resolution corresponds to $w = (1.44 - 1)/0.38 = 1.16$, so under our model his proposed standard error is much too large. The reason for this was indicated above in terms of the implicit variance functions for the overdispersion.

For evaluating the approximation of our Eq (1), we simulated 10,000 trials under the hypothesis fit parameters, finding that 2.8% of the r_ψ-values were greater than the observed value of $r_\psi = 2.19$,

compared to the 2.4% predicted by $\Phi(-r_\psi^*)$ as in (1). For comparison we have that from the first-order theory $\Phi(-r_\psi) = 0.014$.

As noted, the r^* computations using the functions of Figure 3 employ numerical integration to evaluate the likelihood function. Much of the computing time is due to using that method of likelihood evaluation to compute numerical loglikelihood derivatives in the simulation trials. The computing time on a 2.8 GHz laptop is 2 sec without coded gradient and 0.75 sec when using coded gradient function, provided in the R code in the HTML documentation of the package. This difference is amplified when using the more computationally intensive confidence interval routine that inverts the hypothesis test. This is even more important for the bootstrap simulations of the null hypothesis distribution of r,+ which requires many likelihood evaluations even though it does not involve repeated computation of r^*. Nevertheless, numerical results reported above agree to the precision given for the two differentiation methods. As noted above, more important than computing time can be the accuracy of differentiation of a likelihood computed by quadrature.

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 1986-96. In a major advance, Skovgaard (1996, 2001) developed a way of approximating these to second order that involves only computing some loglikelihood-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 $\hat{i}^{-1}\hat{j}$, which is part of the reason we like to think in terms of the Efron-Hinkley ancillary.

An alternative type of approximation, in principle more exact, was developed by Fraser and colleagues. The contrast between these approaches is considered in an Appendix. It was pointed out by Skovgaard (1996) that the two approximations are equivalent when there are no nuisance parameters.

In the Skovgaard approach, it is often unreasonable to compute the required loglikelihood-based covariances 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, since the aim is only estimation of covariances, rather than tail probabilities directly.

Our aim is to approximate the sample-space derivative $\partial^2 l(\psi, \tilde{\lambda}_\psi; \hat{\theta}, a) / (\partial\theta \partial\hat{\theta}^T)$ and $\partial\{l(\hat{\psi}, \hat{\lambda}; \hat{\theta}, a) - l(\psi, \tilde{\lambda}_\psi; \hat{\theta}, a)\} / \partial\hat{\theta}$, that arise in the Jacobian of Section 4. Note that NP_ψ and INF_ψ of (13) can be calculated from those quantities.

The essence of the Skovgaard argument is that, as seen in Fig. 1 of Skovgaard (1985), spaces of fixed ancillary are in a sense "parallel". This parallelism is approximately the same as arises in orthogonal projection, using the observed information inner product, and it is such projection that yields the results,

$$\begin{aligned} \partial^2 l(\psi, \tilde{\lambda}_\psi; \hat{\theta}, a) / \{\partial\theta \partial\hat{\theta}^T) &\doteq \mathrm{cov}_{\hat{\theta}}\{U(\tilde{\theta}), U^T(\hat{\theta})\} \hat{i}^{-1} \hat{j} \\ \partial\{l(\hat{\psi}, \hat{\lambda}; \hat{\theta}, a) - l(\psi, \tilde{\lambda}_\psi; \hat{\theta}, a)\} / \partial\hat{\theta} &\doteq \mathrm{cov}_{\hat{\theta}}\{l(\hat{\theta}) - l(\tilde{\theta}), U(\hat{\theta})\} \hat{i}^{-1} \hat{j} \ , \end{aligned} \tag{14}$$

where the covariances are computed without conditioning on an ancillary. The functions $U(\cdot)$ are ordinary score statistics $U(\theta) = \partial l(\theta; y) / \partial\theta$. The final terms $\hat{i}^{-1}\hat{j}$ 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 moderate deviations, namely $O_P\{\| \hat{\psi} - \psi_0 \| / n^{-1/2}\}$, where ψ_0 is the hypothesis value.

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, e. g. 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. The number of simulation trials for the full exponential family case need be only greater than $\dim(\theta)$, so that matrices involved are positive definite.

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 most clear from the more explicit arguments of Severini (1999). He implicitly presumes that there is an exact ancillary, although it appears it would not be hard to weaken that to utilize a second-order ancillary.

6. DISCRETE DATA AND THE r^* APPROXIMATION

The r^* approximation applies well to discrete data. Unless the distribution of r is highly discrete, taking on with high probability only a few values, one need not be concerned with the complication.

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 $p\{r_\psi(Y) \le r_\psi(\mathrm{y}) : \psi(\theta) = \psi\}$ a continuity correction should be applied. But for discrete settings there are reasons to utilize the mid-*P*, which is the probability of strictly greater evidence plus one-half the probability of the observed result. The term mid-*P* arises from this being the average of the two values where the inequality just given is taken as strong and weak. Due to the discreteness, this mid-*P* has more nearly a uniform distribution under the hypothesis. It is easily seen that the result of employing r^* without any continuity correction provides an approximation to the mid-*P* ; see Pierce and Peters (1999) and Davison, Fraser and Reid (2006).

Considering this more precisely is in principle not complicated. 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 be done in terms of the distribution of r, or in terms of the discrete distribution of *P*-values. Though there are some general formulae for this, summarized by Pierce and Peters (1992), it is typically simpler 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 must conform to this.

So, ignoring the continuity correction leads approximately to the mid-*P*. When the distribution of r is highly discrete, this differs considerably from the usual *P*-value. There are reasons supporting the view that approximating the distribution of r as continuous, without continuity correction, may be preferable. In particular, this may be true when the discreteness arises largely from conditioning, the conditional sample space may be too limited to be useful, or even degenerate. Pierce and Peters (1999) argued that in this case treating the distribution of r as continuous provides a means of approximate conditioning.

The following two examples, though dealing with discrete data, utilize another important issue in regard to the r^* theory that was raised shortly before Example 1 and discussed further in EndNote 1.

EXAMPLE 5. 2 x 2 contingency table

Consider testing independence in the 2x2 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 above. 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. The conditioning on the sufficient statistics for the remaining coordinates is

done automatically in the r^* theory presented here, since as noted above the marginal distribution in Eq. (11) is in this full exponential family setting agrees to third order with the conditional distribution.

```
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 2x2 contingency table

Since 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 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 1-sided *P*-value is 0.0646, and the r^* *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^* *P*-value without continuity correction is 0.0362. This agreement is less precise, but the argument for it given by Pierce and Peters (1999) is more approximate than for the other comparison. Our preference is the r^* *P*-value without continuity correction, but the other numbers are of some interest.

7. DISCUSSION AND CONCLUSION

When the model is a full-rank exponential family, the sample-space derivatives are readily calculated and the simulation in our R routines is not necessary. For the case that the interest parameter is a coordinate of the natural parameter, r^* takes on particularly simple form, as shown in Pierce and Peters (1992). For other interest parameters, Barndorff-Nielsen and Cox (1994) give the formulas in their Example 6.24. The package likelihoodAsy does not attempt to identify these special cases, but anyway

gives the exact results as discussed in Section 5. If the user recognizes the simple form, the number of simulation trials can be reduced to about $\dim(\theta)$, but this is not necessary.

An important aspect of higher-order asymptotics is that inference beyond first order generally depends on more than the likelihood function. To a large extent, the effects of these extra-likelihood aspects are isolated in the *INF* adjustment; see Pierce and Peters (1994), Pierce and Bellio (2006). Since as noted here the *INF* adjustment is usually relatively small, this indicates that even 'exact' inference depends modestly on extra-likelihood considerations. This is particularly important in regard to sequential experiments with data-dependent stopping rules, and in particular sequential clinical trials.

It was noted in Section 1, following items (i) – (iv), that there are two motivations for ancillary conditioning. The one stressed in this paper is that since ancillary conditioning results asymptotically in no loss of information (or power), and generally $\hat{\theta}$ is sufficient conditionally on a suitable ancillary (Skovgaard, 1985), this allows for use of sufficiency principles in considerations of 'optimality'. This is in marked contrast to the Neyman-Pearson theory, where exact optimality is rare, and approximate optimality is not an integral part of the theory as it is for methods of this paper. The motivation for ancillary conditioning more commonly stressed is to render the inference most 'relevant'. A simple way to think of this is that when $\hat{\theta}$ is not sufficient, then the observed information $\hat{j}$ varies around the expected information, the observed information can be seen as more 'relevant'. Efron and Hinkley (1978) provided an excellent discussion of how to directly use the value of $\hat{j}$ for this, and the ancillary conditioning of the present paper resolves this more accurately.

Much of the remainder of this section pertains to the situation where $\dim(\psi) > 1$. Cox and Hinkley (1974, Section 9.3) discuss why there is no uniformly most powerful inference in this case. This is partly why many would agree that it is best when possible, to choose, perhaps in turn, one-dimensional hypotheses as in the present paper. But this is sometimes not attractive, as when testing the null hypothesis in ANOVA or for testing independence in contingency tables.

Skovgaard (2001, Section 5.7) considered extending the theory of this paper to multidimensional interest parameter ψ, proposing a w^* related to the chi-square distribution as r^* is related to the normal. Unfortunately, Skovgaard's proposal based on w^* does not have the distributional accuracy of results presented here for r^* when $\dim(\psi) = 1$. Progress in this direction has been made by Davison, et. al. (2014). Those results are only for full-rank exponential families, for conditioning as considered in EndNote 1, but more general results have been obtained and are likely to appear soon. Because work in this area is rapidly evolving at this time, here we only describe briefly higher order results less

accurate but classical for the case that $\dim(\psi) > 1$: namely the Bartlett adjustment to a chi-squared likelihood ratio test on $\nu = \dim(\psi)$ degrees of freedom. It is well known that if w_ψ denotes the usual first-order chi-squared likelihood ratio test based on n observations, and if we express $E\{w_\psi\} = \nu\{1 + b_\psi / n\} + O(n^{-2})$, then the adjusted statistic $w_\psi^\dagger = w_\psi / \{1 + b_\psi / n\}$ has to second order a chi-square distribution on ν degrees of freedom; see, e.g., Cox and Hinkley (1974, Section 9.3). It is widely acknowledged that in practice the typically best way to assess the bias factor b_ψ is by the parametric bootstrap, e.g. Brazzale, Davison and Reid (2007, Sections 6.3, 7.5).

The *modified profile likelihood* (*MPL*) briefly introduced in Section 4 for the situation where $\dim(\psi) = 1$ also applies to the situation $\dim(\psi) > 1$. It is defined in the notation of Section 4 by

$$L_{MP}(\psi) = L_P(\psi) |\, \tilde{j}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) \,|^{1/2} |\, \partial^2 l(\psi, \tilde{\lambda}_\psi; \hat{\theta}, a) / (\partial\lambda\, \partial\hat{\lambda}^T) \,|^{-1},$$

which can be approximated as in Section 5. Our package likelihoodAsy has provisions for calculating this. The MPL can be useful for Bayesian inference as a pseudo-likelihood, to avoid specifying a prior for the nuisance parameter. For frequentist methods the MPL has the advantage that for $\dim(\psi) > 1$ no method has emerged with the ideal behavior of r^*. The MPL is notably accurate and convenient for logistic regression when conditioning on stratum parameters; see Davison (1988).

Alternatively to L_{MP}, Cox and Reid (1987) proposed an *adjusted profile likelihood (APL),* $L_{AP}(\psi; y) \propto L_P(\psi; y) |\, \tilde{j}_{\lambda\lambda} \,|^{-1/2}$, differing from the *MPL* only by omitting the difficult Jacobian-sample-space derivative term $|\, \partial\hat{\lambda}_\psi / \partial\hat{\lambda} \,| = \{|\, \tilde{j}_{\psi\psi} \,|\, |\, \partial^2 l(\hat{\theta}_\psi) / \partial\lambda\, \partial\hat{\lambda}^T \,|^{-1}\}$. For the *APL* this was dealt with by requiring that the parameters $\{\psi, \lambda\}$ be *orthogonal*, in the sense that $i_{\psi\lambda} = 0$, in which case the omitted term is $|\, \partial^2 l(\hat{\theta}_\psi) / \partial\lambda\, \partial\hat{\lambda}^T \,| = \hat{j} + O_P(n^{-1})$. Unfortunately, though, there are many choices of orthogonal parameters, and hence the *APL* is of limited value. It is precisely this omitted term that renders the *MPL* invariant to the representation of the nuisance parameter. Further, the limitations of the *APL* result from it not being a 'large deviation' result, as is the *MPL*.

Somewhat relatedly, there were various attempts not considered here, usually called 'stable' approximations to r^*, to deal with the difficulty of computing the Jacobian terms indicated at (12) and in Section 5; see Severini (2000), Sect. 7.5. Among the tractable approximations, few if any were very accurate in comparison to the Skovgaard and Fraser methods considered here. This is because even though those other proposals were accurate to second order, they failed to have the protection for large deviations which is more important than the order of approximation, discussed in Section 3 and End-Note 3.

Finally, some have raised the question of the inferential basis for r^*-based inference for reasons other than above, e. g. Reid (2005). Often the r^* inference is couched in terms of the null distribution of r^*, but as emphasized in this paper one can think of the r^* *P*-value with more inferential clarity as a second-order approximation to the distribution of r. Thus the improvement on the ordinary usage of r pertains only to improving on the standard normal approximation used. Pierce and Bellio (2006) emphasized in this respect that the commonly-considered other alternatives to first-order methods, such as modified profile likelihood, provide to second order the same ordering of samples.

ENDNOTES

Note 1. For full-rank exponential families, where the maximum likelihood estimator is sufficient, and when the nuisance parameter comprises a subset of the canonical parameters, it is advantageous and usual to condition on the sufficient statistics associated with those nuisance parameters. For exposition of the Neyman-Pearson approach to this type of inference, see for example, Cox and Hinkley (1974, Section 5). The motivation of this conditioning is to obtain *P*-values whose definition does not depend on the nuisance parameter, pertaining to what are termed *similar tests.*This differs from the ancillary conditioning emphasized in this paper, but Pierce and Peters (1992) and others have shown that the same definition of r_ψ^* derived in Section 4 is appropriate for higher-order conditional inference in this special setting; see also Skovgaard (1987) and Davison (1988). This is somewhat surprising since it does not involve ancillarity, and the derivation of r_ψ^* in the present paper Eq (11) marginalizes over the nuisance parameter estimator $\hat{\lambda}_\psi$, rather than conditioning on it as in the Pierce & Peters (1992) paper. The point is that in this special setting, r_ψ and $\hat{\lambda}_\psi$ are to third order stochastically independent (Jensen, 1986), where he aims to show that in this setting "the primitive procedure [of simply replacing nuisance parameters by estimates] is actually an ideal procedure".

Note 2. 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, Reid and Fraser (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: (a) Skovgaard (1985) and others have shown that under quite general regularity conditions the ancillary $a = \Gamma(\hat{\theta})\hat{i}^{-1}\hat{j}$ of Section 2 meets the requirements (iii) and (iv) of that section, although not uniquely, and (b) this ancillary conforms well to the requirements for the likelihood ratio approximation of Section 3. Efron

and Hinkley (1978) had earlier shown more discursively the value of this ancillary, conjecturing some of the results established by Skovgaard, who showed that this statistic is locally second-order ancillary.

Note 3. The issues of second and third order, i.e. $O(n^{-1})$ 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 equation (7.4) of Severini (2000). The distinction also arises in comparing the approximations to sample-space derivatives due to Fraser and colleagues; e.g. Fraser (2004), and that due to Skovgaard (1996) employed for this paper. For approximating the distribution of r_ψ, as we emphasize 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. (4) 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 and Reid (1987), Severini (2000, Section 7.5.2).

Note 4. 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: e.g. 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 and 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.

Note 5. 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 equations 6.102 – 6.108 of Barndorff-Nielsen and 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.

APPENDIX. AN ALTERNATIVE APPROXIMATION TO THE SAMPLE-SPACE DERIVATIVES

The most widely promoted approach (Davison, 2003, Ch. 12, Brazzale and Davison, 2008, Lozada-Can and Davison, 2010) to computing the sample-space derivatives required for (12) is that due to Fraser and co-workers. Here we use the notation r^* for all of the exact form, and approximate forms computed using either the Fraser or Skovgaard approach. A primary reason we utilize the Skovgaard approach is that, in a sense, it is more general than the Fraser method. For example, in Ex 2 involving dependent AR-1 data, we indicate at the end of this Appendix the special approach employed by Lozada-Can and Davison (2010). Numerical comparison of the Fraser and Skovgaard approaches has been provide by Reid (2003) and by Fraser and Reid (2010). In the examples considered, the numerical difference is quite small and would not be an important basis for choice of the method. These examples are simple and not highly demanding, so there may be others where the choice has greater effect.

In general terms that Fraser method is based on data variations that are locally tangent to the space of fixed-ancillary variation in the sample space, this indeed being precisely the need for the partial derivatives. A somewhat more concrete but equivalent approach to the same end is given in Severini (2000, Section 6.7.2). For independent continuously-distributed observations, there is a relatively simple general formula for the Fraser approach. Since this is a broad class of models, the Fraser approach is not without some generality. For discrete data this must be modified as indicated by Davison, Fraser and Reid (2006). For dependent data more major modifications must be made, as indicated by Lozada-Can and Davison, 2010) for our AR (1) Example 2, and such modifications are generally specific to the nature of the model.

ACKNOWLEDGEMENT

This work 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. Work of both authors was also supported by a grant from the Italian Ministero dell'Istruzione, dell'Universita` e della Ricerca (Prin 2008, Unit of University of Udine). The authors are grateful to Dawn Peters for suggestions that improved the exposition, and to Thomas Severini for answering our technical questions without fail.

REFERENCES

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

Barndorff-Nielsen, O. E. (1984). On conditionality resolution and the likelihood ratio for curved exponential models. *Scand. J. Statist.* **11**, 157-170.

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

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

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

Brazzale, A. R. and Davison, A. C. (2008). Accurate parametric inference for small samples. S*tatist. Science* **23**, 465-484.

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

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

Cox, D. R. (1970) *Analysis of Binary Data.* Chapman and Hall, London.

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

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

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

Cox, D. R. (2006). *Principles of Statistical Inference.* Cambridge Univ. Press, Cambridge.

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

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

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

Davison, A. C., Fraser, D. A. S., Reid, N. and Sartori, N. (2014). Accurate directional inference for vector parameters in linear exponential families. *J. Amer. Statist. Assn.* ***109****, 302-314.*

Davison, A. C., Hinkley, D. V. and Young, G. A. (2003). Recent developments in bootstrap methodology. *Statist. Sci.* **18**, 141-147.

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

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

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

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457-482.

Efron, B. (1998). R. A. Fisher in the 21^{st} century: Invited paper presented at the 1996 Fisher Lecture. *Statistical Science* **13**, 95-122.

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

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

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. A* **144**, 295-307.

Fraser, D. A. S. (1991). Likelihood to significance. *J. Amer. Statist. Assn.* **86**, 258-265.

Fraser, D. A. S. (2004). The 1990 Fisher Lecture: Ancillaries and conditional inference. *Statist. Sci.* **19**, 333-369.

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

Fraser, D. A. S. and Reid, N. (2010). Mean likelihood and higher-order approximations. *Biometrika* **97**, 159-170.

Ghosh, M., Reid, N. and 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**, 567-572

Johansen, S. (1995). The role of ancillarity in inference for non-stationary variables. *The Economij Journal* **105**, 302-320.

Kalbfleisch , J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data. 2nd Ed.* John Wiley and Sons, New York.

Lawless, J. F. (1973). Conditional versus unconditional confidence intervals for parameters of the Weibull distribution. *J. Amer. Statist. Assn.* ***68***. 665-669.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data, Second Edition.* John Wiley & Sons. Hoboken, New Jersey.

Lozada-Can, C. and Davison, A. C. (2010). Three examples of accurate likelihood inference. *The Amer. Statist.* **64** , 131-139.

McCullagh, P. (1984). Local sufficiency. *Biometrika* ***71****,* 233-244.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear and Mixed Models, 2nd Ed*. John Wiley and Sons, New York.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective.* World Scientific Publishing Co. , Singapore.

Pierce, D. A. and Bellio, R. (2006). Effects of the reference set on frequentist inferences. *Biometrika* ***93****,* 425-438.

Pierce, D. A. and Bellio, R. (2015). Beyond first-order asymptotics for Cox regression. *Bernoulli* ***21****, 401-419.*

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

Pierce, D. A. and Peters, D. (1994). Higher-order asymptotics and the likelihood principle: One-parameter models. *Biometrika* **81**, 1-10.

Pierce, D. A. and Peters, D. (1999). Improving on exact tests by approximate conditioning. *Biometrika* ***86****, 265-277.*

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. *http://www.R-project.org*

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statist. Science* **3**, 213-227.

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

Reid, N. (2003). The 2000 Wald Lectures: Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695-1731.

Reid, N. (2005). Theoretical statistics and asymptotics. Chapter in *A Celebration of Statistics: Papers in Honour of D. R. Cox*. Eds. A. C. Davison, Y. Dodge and N. Wermuth. 73-85, Oxford Univ. Press, Oxford.

Severini, T. A. (1990). Conditional properties of likelihood-based significance tests. *Biometrika **77**, 343-352.*

Severini, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika* **86,** 235-247.

Severini, T. A. (2000). *Likelihood Methods in Statistics.* Oxford Univ. Press, Oxford.

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

Skovgaard, I. M. (1987). Saddlepoint approximations for conditional distributions. *J. Appl. Prob. **24**, 875-877.*

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

Skovgaard, I. M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28**, 3-32.

StataCorp (2015). *Stata Statistical Software: Release 14*. College Station, TX

Sweeting, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1-23.

Young, G. A. (2009). Routes to higher-order accuracy in parametric inference. *Australian & New Zealand Journal of Statistics* **51**, 115-126

Young, G.A. and Smith, R. L. (2005). *Essentials of Statistical Inference.* Cambridge Univ. Press, Cambridge.