The Effect of Choice of Reference Set on Frequency Inferences

By Donald A. Pierce

Radiation Effects Research Foundation, 5-2 Hijiyama Park,
Hiroshima 732-0815, Japan
pierce@rerf.or.jp

And Ruggero Bellio

Department of Statistics, University of Udine, Via Treppo 18, Udine 33100, Italy
ruggero.bellio@dss.uniud.it

SUMMARY

It is well known that a given dataset can be thought of within various reference sets, with no effect on the likelihood function but leading to different frequency inferences. Here we focus on the effect of censoring models for response-time data and stopping rules for sequential experiments, although our results have more general implications. Since the choice of reference set does not affect first-order likelihood-based inferences, higher-order likelihood asymptotics providing explicit corrections to these are ideally suited to assessment of reference-set effects. It has been noted that there are two aspects of higher-order corrections to first-order likelihood methods: (i) that involving effects of fitting nuisance parameters and leading to the modified profile likelihood, and (ii) another part pertaining to limitation in adjusted information. We show that correction (i) is largely independent of the reference set, whereas (ii) is what reflects that choice. In particular, to second order (i) is not affected by either censoring models or stopping rules, whereas (ii) is affected by stopping rules but not censoring models. These structural results bear on issues regarding likelihood-Bayesian *vs* frequency inference, clarifying that modified profile likelihood is appropriate for both modes of inference through its independence of the reference set, and it is only in regard to adjustment (ii) that they should differ in this respect. Two side issues are addressed, one being the applicability of higher-order likelihood asymptotics to sequential settings. The other pertains to the meaning of ideal inferences for evaluating reference set effects, since exact methods will not exist generally enough for this purpose. In this regard it is shown that the ordinary signed likelihood ratio statistic, its usual higher-order modifications, and the version arising from integrating out nuisance parameters using any smooth prior, all provide to second order the same ordering of datasets according to evidence against the hypothesis. Thus in a strong sense all of these test statistics are second-order equivalent, even though to this order they have different limiting distributions.

Some key words: censoring model, higher-order asymptotics, likelihood asymptotics, likelihood principle, modified directed deviance, modified profile likelihood, sequential experiments.

1. INTRODUCTION

It is well known that a given dataset can be thought of within various reference sets, *i.e.* hypothetical repetitions of an experiment, leading to different frequency inferences. We are interested only in choices of reference set resulting in the same likelihood function for the data at hand. See, for example, Cox & Hinkley (1974, § 2.3, 2.4). There are various reasons for interest in the general matter, including: (i) foundational issues involving the likelihood principle or contrast of frequency and Bayesian inference, (ii) both theoretical and practical unattractiveness of inferential dependence on models for censored or otherwise incomplete data, and (iii) further insights into how inferences should depend on stopping rules in sequential experiments. Our aim is to clarify the structural dependence of frequency inference on the choice of reference set, with specific emphasis on (ii) and (iii) but also adding clarification to (i).

Our interest is in directional inference regarding a scalar parametric function $\psi(\theta)$ in multiparameter problems, couched in terms of significance tests but largely with the aim of inverting these for confidence intervals. We consider only distributions under the hypothesis, as required for these aims. Almost never is there an 'exact' test for hypotheses of this nature within each of two reference sets, since this would require that in both settings the problem is either a full-rank exponential family or a transformation model. We will thus refer to 'ideal' rather than 'exact' inference, which involves asymptotic considerations dealt with in § 2.

Sequential settings raise issues regarding the validity of usual first order asymptotic methods, considered in § 4, and for the meantime we note that our general considerations apply for stopping rules where first-order methods are valid. The asymptotic index is generally an information matrix scaling factor that we usually denote informally by n; often this is the number of independent observations but for sequential settings the index is essentially $E(n)$. All results are for so-called moderate deviations, basically where $\hat{\theta}-\theta=O_p(n^{-1/2})$, and thus little distinction is required between $O_p(\cdot)$ and $O(\cdot)$. When we say that a relation holds to $O(n^{-k/2})$, *i.e.* to k^{th} order, we mean when ignoring terms of $O(n^{-k/2})$ in asymptotic expansions. Usually the reference quantity will be $O(1)$ and if not we will strive to make clear whether this refers to absolute or relative error.

Pierce & Peters (1994) noted that modern higher-order asymptotics, in the form lately called likelihood asymptotics, provides for incisive quantitative investigation of reference set issues. Likelihood asymptotics express ideal frequency directional *P*-values explicitly in terms of

(a) the contribution from the directional likelihood ratio statistic, which is in our framework both independent of the choice of reference set and first-order standard normal for any such choice; and

(b) a higher-order adjustment depending on more than the likelihood function and thus capturing the effect of the reference set, along with other aspects of moving from first-order likelihood inference to more precise *P*-values.

The adjustment (b) generally affects *P*-values by $O(n^{-1/2})$, yielding accuracy to at least $O(n^{-1})$. The main points to be made are that sometimes the effect of the reference set on that adjustment is of $O(n^{-1})$, and that usually the effect on a major aspect of it is of that order. It might be said that effects of that magnitude are nil since it seems that, generally enough for purposes of comparing reference sets, ideal inferences may only be defined to $O(n^{-1})$.

Barndorff-Nielsen & Cox (1984 and 1994, § 7.5), and Sweeting (2001) used likelihood asymptotics to study reference set effects, considering both censoring and sequential settings. However, they only considered 2-sided *P*-values, where directional effects often cancel to second order, and thus some of their statements may superficially seem contradictory to ours.

For one-parameter models, Pierce & Peters (1994) used the adjustment (b) referred to above to investigate the magnitude of the $n^{-1/2}$ term of the dependence of *P*-values on stopping rules in highly stylized one-parameter settings. In other considerations they failed to realize that, as shown here, the effect of censoring model on the term (b) is of $O(n^{-1})$. To an extent this is simply because potential censoring times, when they are all known, are ancillary statistics, and since ideal inference should be conditional on ancillaries it will not depend on the censoring model. However, potential but unrealized censoring times are often not known, and hence a more general argument is needed. Censoring is an instance of missing or coarsened data, and results here will to some extent apply to these more general settings.

Pierce & Peters (1994) considered only in passing settings with nuisance parameters, making a conjecture that is clarified here. In such settings (b) comprises two parts: one reflecting effects of fitting nuisance parameters and the other pertaining to moving from likelihood-based inference to frequency *P*-values. We show here that the former of these parts depends to at most $O(n^{-1})$ on stopping rules for sequential experiments, where these are considered much more generally than by Pierce & Peters. This part of the higher-order adjustment provides what is referred to as the modified profile likelihood and thus we see that, for the two primary settings considered, modified profile likelihood indeed has the likelihood-like character of being largely free of the reference set.

Before turning to our main results we give some background material. There is by now a vast literature on asymptotic theory related to this paper, and we refer only to papers very directly pertinent to our developments. We largely follow the development due to Barndorff-Nielsen, noting that Fraser and co-workers have developed a somewhat different approach; see for example Fraser, Reid & Wu (1999), Fraser (2003). Surveys of relevant literature are given by Reid (1996, 2004) and Skovgaard (2001), and useful treatments that include introductory matters are given by Pace & Salvan (1997) and Severini (2000).

2. BACKGROUND ON HIGHER-ORDER ASYMPTOTICS

Ideal inference

We denote by y the entire dataset, whose fully-parametric distribution for a given reference set is governed by a multidimensional parameter θ. Consider testing an hypothesis $\psi(\theta) = \psi$ versus one-sided alternatives, where ψ is a scalar function. Ideal inferences, and the higher-order asymptotics of interest here, begin with the directed deviance, *i.e.* the signed square root of the asymptotically χ_1^2 generalized likelihood ratio statistic,

$$r_\psi(y) = \mathrm{sgn}(\hat{\psi} - \psi)\left[2\left\{l(\hat{\theta}; y) - l(\hat{\theta}_\psi; y)\right\}\right]^{1/2}, \qquad (1)$$

where $l(\theta; y)$ is the loglikelihood function, $\hat{\theta}$ and $\hat{\psi}$ are maximum likelihood estimators, and $\hat{\theta}_\psi = (\psi, \hat{\nu}_\psi)$ is the constrained maximum likelihood estimator under the hypothesis. In that last notation, the nuisance parameter is denoted by ν and its constrained estimator by $\hat{\nu}_\psi$. Note that $l(\hat{\theta}_\psi; y)$ is the log profile likelihood and $l(\hat{\theta}; y)$ is its maximum value. It is useful to have the subscript ψ explicitly denoting both the interest parameter and value of the hypothesis being tested. The value of r_ψ does not depend on the reference set, and quite generally this quantity has to first order a standard normal distribution regardless of the reference set.

As already noted, reference set effects must be considered in terms of ideal, rather than exact, frequency inference. Fundamental issues of inference involve only an ordering of datasets y according to evidence against the hypothesis. That is, given such an ordering the evaluation of a *P*-value involves only distributional calculations unrelated to theory of inference. In spite of well-known limitations of using the standard normal approximation to the distribution of $r_\psi(y)$, there are strong reasons for why this statistic provides the ideal ordering of datasets. It is shown in Appendix 1 that the ordering of datasets provided by $r_\psi(y)$ is to second order not altered by employing usual modifications of this statistic to improve its standard normal approximation, nor by using a one-parameter version of $r_\psi(y)$ based on any of the usual pseudo-likelihoods for ψ. This includes the usual modified or adjusted profile likelihoods and also the likelihood arising from integrating out the nuisance parameter according to any smooth prior distribution. These results suggest that to second order the ideal *P*-value for observed data y_{obs} is of form $pr\{r_\psi(y) \leq r_\psi(y_{obs});\psi,\nu\}$, which to this order does not depend on ν. The semi-Bayesian result noted above is to us a particularly compelling part of the justification for this. There are various reasons for why it is unlikely that there is a general theory of frequency inference to third order and thus we refer to the method just given as ideal, rather than ideal to second order. There are further comments on third-order matters below.

We now turn to the usual approximation of such *P*-values arising in likelihood asymptotics. We note that there is basically nothing in the derivation of the following results excluding their applicability to sequential experiments, provided that the stopping rule is such that $r_\psi(y)$ has to first order a standard normal distribution. We return to such issues in § 4.

Approximating the distribution of $r_\psi(y)$

Details and further references regarding the following are provided in Appendix 2. Barndorff-Nielsen (1986; see also § 6.6 of Barndorff-Nielsen & Cox, 1994) proposed an adjusted version of the signed likelihood ratio statistic of form explained below and further in Appendix 2,

$$r_\psi^*(y) = r_\psi(y) + \{r_\psi(y)\}^{-1}\log\{u_\psi(y)/r_\psi(y)\}, \tag{2}$$

such that to second order and for observed data y_{obs},

$$pr\left\{r_\psi(y) \leq r_\psi(y_{obs});\psi,\nu\right\} \triangleq \Phi\left\{r_\psi^*(y_{obs})\right\} \tag{3}$$

with Φ the standard normal distribution function. An implication of (3) is that to second order the left side depends only on ψ. The adjustment to r_ψ is $O(n^{-1/2})$ and as discussed in the Introduction it subsumes the aspect of inference depending on the reference set. Note that (3) is valid if and only if $r_\psi^*(y)$ is standard normal to second order. In fact this latter condition generally holds to third order, and focus on that has rather obscured the second-order relation (3), even though obtaining this was the original motivation for (2). The relation (3) cannot be generally valid to third order, since to that order $r_\psi^*(y)$ is not a function of $r_\psi(y)$; see comments at the end of Appendix 2.

It is in principle possible to compute the left side of (3) to third order without knowledge of the true ν, since DiCiccio, Martin & Stern (2001) have shown that to this order $pr\left\{r_\psi(y) \le r_\psi(y_{obs});\psi,\hat{\nu}_\psi\right\} \triangleq pr\left\{r_\psi(y) \le r_\psi(y_{obs});\psi,\nu\right\}$. Evaluating the left side to this accuracy would usually involve simulation. However, as noted by those authors, there seems not to be a satisfactory inferential basis for *P*-values computed to that order, which is part of the reason we refer to second-order theory as ideal.

In related theory Barndorff-Nielsen (1983), see also § 8.2 of Barndorff-Nielsen & Cox (1994), proposed a modified profile likelihood function $L_{MP}(\psi) = M(\psi)L_P(\psi)$ aiming to reduce undesirable effects of fitting nuisance parameters. Following development by Pierce & Peters (1992) for full-rank exponential families, as generalized in Barndorff-Nielsen & Cox (1994), (2) can usefully be expressed as

$$r_\psi^*(y) = r_\psi(y) + NP_\psi(y) + INF_\psi(y) \qquad (4)$$

in such a way that $\exp(-r_\psi\, NP_\psi) = M(\psi)$. Thus $L_{MP}(\psi) = \exp(-r_\psi\, NP_\psi - r_\psi^2/2)$ and hence to second order $L_{MP}(\psi) = \exp\{-(r_\psi + NP_\psi)^2/2\}$. For these reasons NP_ψ is referred to as the nuisance parameter adjustment. The remaining adjustment INF_ψ pertains more specifically to moving from likelihood to frequency inference, and is called the information adjustment since it is only substantial when the adjusted information for ψ is small. Thus in practice the INF adjustment is often small or negligible, but when there are several nuisance parameters the NP adjustment can be substantial even when the adjusted information for ψ is large. This is exemplified in Example 1 to follow, with many other examples given in Pierce & Peters (1992), § 6.6 and § 8.2 of Barndorff-Nielsen & Cox (1994), Sartori, et al. (1999), Chapters 7–9 of Severini (2000), Sartori (2003) and elsewhere.

Results here are that usually, but perhaps not always, the NP part of the adjustment depends on the choice of reference set only to $O(n^{-1})$, this being one order smaller than the magnitude of the adjustment itself. Thus the modified profile likelihood usually has in this respect the character of an ordinary likelihood function.

Usually the INF adjustment reflects the first-order effect of the choice of reference set along with the effect of other limitations in adjusted information. But in certain settings, notably those involving censored data, the effect of the reference set on INF is also $O(n^{-1})$.

Except for full-rank exponential families and transformation (*e.g.* location-scale) models, the adjustments referred to above are difficult to compute since they involve differentiation of the loglikelihood with respect to parameter estimates while holding fixed an approximate ancillary statistic, so-called sample-space derivatives. Various approximations to these with resulting error $O(n^{-1})$ for the adjustments have been developed, but it was a major advance when Skovgaard (1996) developed the general and fundamental method used here. This is detailed in Appendix 2 and partially indicated in the following section. The error in this approximation is not only $O(n^{-1})$ but is proportional to the statistical curvature of the model, which in practice is usually quite small. We will make no notational distinction for quantities above to reflect use of the Skovgaard approximation.

3. CENSORED DATA

Although we comment later on more general results, the argument considered in detail is restricted to survival times with literally independent censoring. That is, the censored observations t_i are the minimum of independent response times and censoring times that are either fixed or mutually independent random variables also independent of response times. We show that for this setting ideal inference from a given dataset does not depend on further specification of the censoring model. In particular, inference does not depend whether censoring times are fixed or random, or on unobserved random censoring times. As usual, the data are represented as observations (t_i, c_i), where c_i is an indicator of censoring, and the usual contributions to the likelihood of the form $f_i(t_i;\theta)^{1-c_i}\, pr(T_i > t_i;\theta)^{c_i}$ are for our setting stochastically independent. This would not always be true for more general censoring models.

The conditions required for our argument can be stated rather more generally, for broader applications to partially observed data, as follows. For some class of reference sets compatible with the observed data, let there be a representation of the loglikelihood as $l = \sum_{i=1}^{n} l_i$ not depending on the reference set, such that the contributions $\{l_1, l_{2,} \cdots\}$ are stochastically independent. Under these conditions the adjustments NP and INF are to second order independent of the reference sets considered. Our argument involves consideration of $\text{cov}(U, J)$, where U and $-J$ are first and second derivatives of the loglikelihood. Expansions given in the following paragraph show that to second order the sample-space derivatives required for the higher-order adjustments depend on the reference set only through $\text{cov}(U, J)$, evaluated at $\theta = \hat{\theta}$. Further, we see that approximation of this covariance with first-order relative error results in second-order approximation to the sample-space derivatives. Under the above assumptions this can be done independently of the reference set, using the sample mean from the observed data of quantities $(U_i J_i)$, where the U_i and J_i correspond to the loglikelihood contributions.

Although laying out the remaining argument in full detail depends on matters addressed in Appendix 2, the following makes it sufficiently clear. Skovgaard's second-order approximation to the sample-space derivative involved in the NP adjustment is

$$\partial^2 l(\hat{\theta}_\psi; y) / \{\partial\theta \partial\hat{\theta}\} \doteq \text{cov}_{\hat{\theta}} \left\{ U(\hat{\theta}_\psi), U(\hat{\theta}) \right\} \hat{\imath}^{-1} \hat{\jmath} \ , \tag{5}$$

where the covariance of the loglikelihood derivatives at two parameter values is evaluated before substituting the parameter estimates, and where $\hat{\jmath}$ and $\hat{\imath}$ are the observed and expected information evaluated at $\hat{\theta}$. When the term $U(\hat{\theta}_\psi)$ is expanded in $\hat{\theta}_\psi - \hat{\theta}$ the leading term of the covariance in (5) is $\hat{\imath}$, leading to the second-order approximation

$$n^{-1} \partial^2 l(\hat{\theta}_\psi; y) / \{\partial\theta \partial\hat{\theta}\} = n^{-1} \hat{\jmath} + (\hat{\theta}_\psi - \hat{\theta}) n^{-1} \text{cov}(U, J) \hat{\imath}^{-1} \hat{\jmath} + O(n^{-1}) \ ,$$

where the covariance is evaluated at $\hat{\theta}$, and we take some innocuous notational liberties regarding that 3-dimensional array. The $O(1)$ leading term in this expansion does not depend on the reference set, and the second term, which does depend on the reference set through the expectations involved, is $O(n^{-1/2})$. The term $\hat{\imath}^{-1} \hat{\jmath}$ is $I + O(n^{-1/2})$ and can be omitted. Thus approximation of $n^{-1} \text{cov}(U, J) = O(1)$ with first-order error leads to second-order approximation of the sample space derivative. Results develop similarly for the sample-space derivative providing the *INF* adjustment, where the leading term is $n^{-1/2}\hat{j}(\hat{\theta}_{\psi}-\hat{\theta})$ and the second term involves $n^{-1}\,\mathrm{cov}(U,J)$.

Although this establishes the main result of this section, we note that some elaboration on the argument leads to an observation by Severini (1999) that could be used more directly to obtain the end results. He showed that second-order error is introduced in (5) by approximating the quantities $\mathrm{cov}_{\hat{\theta}}\left\{U(\hat{\theta}_{\psi}),U(\hat{\theta})\right\}$ and $\hat{i}$ by empirical covariances of scores computed from the observed contributions to the likelihood. There is a corresponding result for the other sample-space derivative. Thus his approximation depends on the contributions to the loglikelihood but not on the censoring model.

It appears that, in a sense, our main result remains true for general censoring models but since we have not pursued this in detail, further comment on this is relegated to § 5.

Example 1.

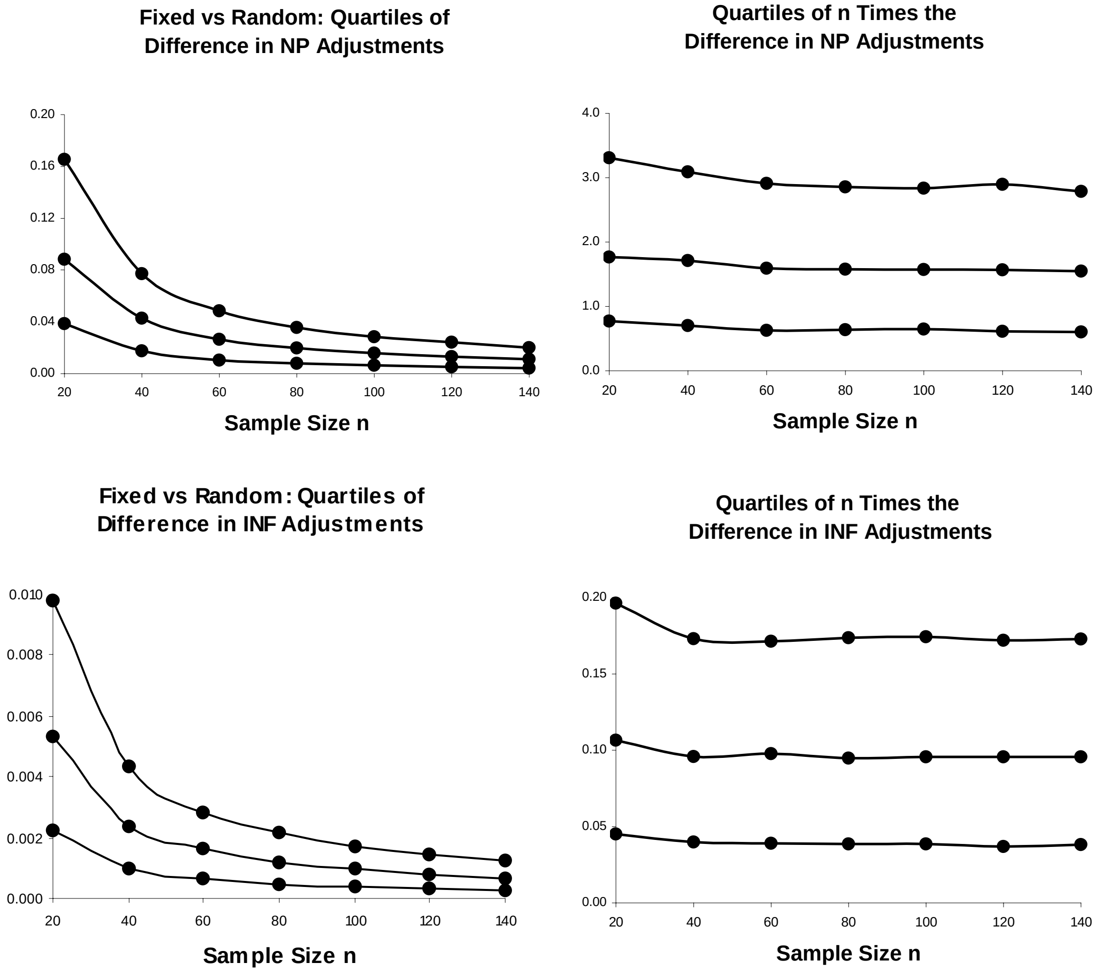
We consider Weibull regression with fixed and random censoring, where interest is on the shape parameter. This example is useful in that the nuisance parameter adjustment is large when there are several covariables, even for fairly large samples. Employing simulation we verify numerically that both the nuisance parameter and information adjustments agree to second order for the two reference sets. The covariances required for the Skovgaard approximation are rather intractable for the regression setting, and we approximate these using a very large number of Monte Carlo trials for each observed dataset in the main simulation. In particular, our simulation used 10,000 datasets with 5,000 Monte Carlo trials for each. In order to carry this out simply for fixed censoring times, we assume that as is often the case, the potential censoring times are all known.

The assumed model has hazard function $\lambda_i(t;\beta,\psi)=\exp(z_i\beta)\,t^{\psi}$, where z_i comprises a constant term and 5 Gaussian covariables, with results given for testing $\psi=1$. Censoring times were generated as exponential variates scaled so that there is 20% chance of censoring, where these were either treated as fixed or random in computing the higher-order adjustments. The approximation (3) is quite accurate in this example, when computed under either censoring model. For true tail probabilities in the range 0.01–0.20 the average absolute value of the relative error is 10% for $n = 60$, whereas this error is 120% when using the unadjusted r_ψ. For further perspective we first give some general idea of the magnitude of higher-order adjustments. Writing $r^* = r + NP + INF$, Table 1 indicates quartiles of each adjustment for fixed-time censoring, which depend minimally on the censoring model. The *NP* adjustment is substantial even for $n = 140$.

Table 1. Description of adjustments

Sample size	Quartiles of –*NP*	Quartiles of –*INF*
20	(1.05, 1.17, 1.29)	(0.10, 0.11, 0.13)
60	(0.54, 0.59, 0.63)	(0.05, 0.06, 0.06)
100	(0.41, 0.44, 0.47)	(0.04, 0.04, 0.05)
140	(0.35, 0.37, 0.39)	(0.03, 0.04, 0.04)

Figure 1 indicates the reference set effect on each of the *NP* and *INF* adjustments. For each adjustment, shown are quartiles of the differences in the adjustments between the two reference sets. It is seen that, in line with our theoretical results, the differences are $O(n^{-1})$ for both adjustments.



Fixed vs Random: Quartiles of Difference in NP Adjustments

Quartiles of n Times the Difference in NP Adjustments

Fixed vs Random: Quartiles of Difference in INF Adjustments

Quartiles of n Times the Difference in INF Adjustments

Figure 1. Quartiles of the distribution of differences in the adjustment terms for fixed and random censoring.

4. SEQUENTIAL EXPERIMENTS

General considerations

Pierce & Peters (1994) considered from the view of this paper a certain class of sequential models for one-parameter settings, exemplified by binomial vs negative binomial, or Gaussian vs inverse Gaussian, sampling. Here we want to introduce nuisance parameters, and consider more general sequential settings including essentially what arises in sequential clinical trials. Our considerations apply whenever there are underlying independent observations $y_1, y_2, \cdots$, with only the first *n* of these observed according to some data-dependent stopping rule. There are some tensions discussed below between classical theory of sequential inference and what is required to compute *P*-values and confidence intervals. Our view, basically compatible with that of many others, is that inferences should remain to be based on the distribution of $r_\psi(y)$. Although results of § 2 for approximating this apply fairly generally to sequential settings, we give some attention to required conditions for this. Our main result is that when that theory does apply, then the *NP* adjustment and modified profile likelihood quite generally depend to second order on the stopping rule, and hence its effect is reflected by the *INF* adjustment.

In the classical theory of sequential analysis, and much of recent considerations, some portion of the stopping boundary is a pre-specified 'rejection region'. Whereas in fixed sample size settings the relation between prespecified α-level rejection rules and post-data *P*-values is relatively simple, this is not the case when the rejection rule corresponds to a portion of the stopping boundary. The difficulty, and a resolution of it, was perhaps first articulated by Armitage (1957). There is by now fairly general agreement that, as he suggested, upon reaching the stopping boundary, *P*-values and confidence intervals should be based on an ordering of points on the entire stopping boundary according to evidence against the hypothesis; see for example Siegmund (1978), Rosner & Tsiatis (1988), Whitehead (1999), and Cook (2002). Various orderings have been proposed, some of which are rather arbitrary or intuitively-based, but we maintain that the general considerations earlier in this paper ordinarily apply, so that the ideal ordering is based on the value of $r_\psi(y)$. This notion is reasonably consistent with results of specific investigations regarding preferred orderings in papers cited above. Computing *P*-values in this manner is equivalent to basing them on the distribution of $r_\psi(y)$ over the stopping boundary, which of course depends on the stopping rule. We aim to use results in § 2 to understand this dependence.

A common view might be that in a final reduction of the data to (r_ψ, n) the value of n is important to the inference. The value of n is certainly informative, but for datasets on the stopping boundary it is often substantially predictable from $\hat{\theta}$. Thus it is inferentially more appropriate to consider the data summary as (r_ψ, a_n), where

$a_n = \{n - E_{\hat{\theta}}(n)\} / SD_{\hat{\theta}}(n)$. It is a standard result that under suitable regularity conditions a_n is to first order standard normal. When asymptotics considered here apply, then a_n is asymptotically ancillary and to second order r_ψ is stochastically independent of ancillaries, so inference may be based on the marginal distribution of r_ψ. In one-parameter settings the statistic a_n can be closely related to the Efron-Hinkley ancillary (Efron & Hinkley, 1978), but generally a_n comprises only part of the ancillary information in what we have written as a.

The underlying observations $y_1, y_{2,} \cdots$ will now be taken as independent and identically distributed. Comparative inference and regression aspects can to an extent be included by thinking of each y_i as a vector with an associated covariable not depending on i. The identically distributed condition could be avoided through conditions on the sequence of covariables. The asymptotic index in likelihood asymptotics is a scaling factor of the Fisher information, which in our setting is proportional to $E_\theta(n)$. As for stopping rules, our main results require only the usual condition that stopping at trial n depends only on data $y_1, y_{2,} \cdots, y_n$. It is well known that under this condition the likelihood function does not depend on the stopping rule, see for example, Ex. 2.34 of Cox & Hinkley (1974).

We focus in our examples on stopping rules of form $| r_\psi(y_1, \cdots, y_n) | > \eta c_n$, for some hypothesized value ψ and given sequence c_n, with the additional proviso that $n \leq N$ for some fixed N. This is a generalization of rules based on repeated significance tests, see for example Armitage (1991). For asymptotics η and N are taken to increase together. Important modifications of this form to which our results apply arise when the parametric function of inferential interest is different from that involved in the stopping rule, as considered by Whitehead, Todd & Hall (2000), and when the stopping rule pertains to the precision of the inference, as considered by Grambsch (1983).

Even usual first-order asymptotics can fail in sequential settings, and conditions for them to hold were established by Anscombe (1952). Anscombe's Theorem shows that if $r_\psi(y)$ is to first order standard normal for fixed sample sizes, then this remains true for sequential sampling under two conditions: basically that the coefficient of variation of the stopping time *n* approaches zero, and that the distribution of $r_{\psi}(y)$ is asymptotically suitably continuous. Grambsch (1983) provides a statement of Anscombe's Theorem, giving arguments and results tangential to ours. The condition on the stopping time is largely a formality, and for applications the critical issue is the continuity condition. As Anscombe showed, the discrete time aspect does not interfere with this. However, the nature of the stopping boundary can affect that continuity in complicated ways. As a simple but important example, consider stopping at the smallest n where $|r_{\psi}| \geq c$ or $n = N$, where asymptotically c and N increase together. Then clearly Anscombe's Theorem can only apply to the distribution of r_{ψ} when $pr(n < N)$ is asymptotically negligible.

Higher-order asymptotics

As noted by Barndorff-Nielsen & Cox (1994, Sect 7.5) and Coad & Woodroofe (1996) issues can arise regarding validity of the results of § 2 for sequential settings. We will indicate below that a necessary and sufficient condition for this validity is that the distribution of $r_{\psi}(y)$ is to first order standard normal, *i.e.* that Anscombe's Theorem applies.

It is well known that when Anscombe's Theorem does apply, the convergence may be slow and higher-order considerations are important. Others have utilized more classical asymptotics than here, for example Woodroofe & Keener (1987). Woodroofe (1992) and Coad & Woodroofe (1996) take an approach more closely related ours, which is considered later in Example 2. Although we indicate that higher-order likelihood asymptotics fairly generally apply for stopping rules such that first-order methods are valid, our aims are in a sense less ambitious than usual. Sequential settings may sometimes stress second-order theory, and we recommend that ordinarily the actual computation of *P*-values be done by direct simulation of the distribution of r_{ψ}, using for the nuisance parameter the estimate $\hat{\nu}_{\psi}$. Our aim is less to compute *P*-values than to clarify in moderate generality structural aspects of the effect of stopping rules on the second-order distribution of r_{ψ}, and thus their usual effect on ideal inference. The direct simulation approach was suggested by DiCiccio, Martin & Stern (2001), although they did not have in mind sequential settings.

Arguments underlying results of § 2 require, strictly speaking, that all relevant quantities have densities with respect to Lebesgue measure, or some other suitable measure not depending on n, which will not be the case for discrete-time sequential settings. As in other discrete-data settings, the best resolution of this is to think of the development in terms of a continuous idealization, then verifying either theoretically or numerically that results apply to the actual settings. We assume that such idealizations are made, and consider only more fundamental issues regarding the form of stopping rules.

The substantive usual condition required for higher-order likelihood asymptotics to apply is that $p(\hat{\theta} \mid a; \theta)$ is to first order normal with something akin to a second-order Edgeworth expansion. In particular, there is no assumption of independent observations. However, we do not want to make assumptions regarding $p(\hat{\theta} \mid a; \theta)$ since there are available no results regarding the effect of stopping rules on this. To deal with the needs in more familiar terms, we indicate in Appendix 3 that what is required for results of § 2 is only that the stopping rule is such that r_{ψ} is standard normal to first order and that its density allows something akin to a second-order Edgeworth expansion. The argument is not specialized to sequential settings and clarifies in general the role of conditioning on ancillaries in higher-order likelihood asymptotics. That is, to second order conditioning on ancillaries is not ultimately required since r_{ψ} is independent of these, and it is interesting to see just where in the development of higher-order asymptotics this conditioning can be dropped.

Argument for main result

We now show that to second order the NP adjustment and modified profile likelihood function $L_{MP}(\psi)$ do not depend on the stopping rule. Although $L_{MP}(\psi)$ and this result have nothing to do with representation of the nuisance parameter, the argument involves specifications where ν is orthogonal to ψ. The key to the argument is showing that such orthogonality does not depend on the stopping rule. We also show that when there is a choice of nuisance parameter meeting the stronger condition that $\hat{\nu}_{\psi} \equiv \hat{\nu}$, the NP adjustment and modified profile likelihood are exactly independent of the stopping rule.

For the general argument we rely on our formulation where $y_1, y_2, \cdots$ are identically distributed, although this is not totally necessary. Write $i_1(\theta)$ for the expected information corresponding to a single observation y_i , noting that in the fixed-n setting the information in the sample is $i(\theta) = n\ i_1(\theta)$. Then for the sequential setting it follows from Wald's Fundamental Identity that $i(\theta) = E_\theta(n)\ i_1(\theta)$; see in particular § 7c.2 (v) of Rao (1973). A more modern view of this is that $\partial^2 l_n / (\partial\theta)^2$ less its compensator, where l_n is the loglikelihood for the first n observations, is a martingale relative to the sequence $y_1, y_2, \cdots$, and thus the expectation of $\partial^2 l_n / (\partial\theta)^2 - n i_1(\theta)$ is zero regardless of the stopping rule. So we see that parameters orthogonal in the fixed-n setting remain orthogonal in the sequential setting.

As is well known and noted in Appendix 2, $L_{MP}(\psi)$ agrees to second order with the Cox-Reid approximate conditional likelihood $L_{AC}(\psi)$ for any choice of orthogonal nuisance parameter, which depends only on the likelihood function. Since the orthogonal parameter can be defined independently of the stopping rule, this means that to second order $L_{MP}(\psi)$, and hence the NP adjustment, do not depend on the stopping rule. The limitation that $L_{AC}(\psi)$ depends on the specific choice of orthogonal nuisance parameter is not very serious for the present argument, since it holds for every such choice and one can expect that usually for some choice the approximation of $L_{MP}(\psi)$ by $L_{AC}(\psi)$ is quite good.

For the further result suppose there is a choice of nuisance parameter such that $\hat{\nu}_\psi \equiv \hat{\nu}$, which is the case in any full-rank exponential family when ψ is a canonical parameter. In particular, ν can then be taken as the complementary mean parameter; see for example § 2.9 of Barndorff-Nielsen & Cox (1994). It is well known that when $\hat{\nu}_\psi \equiv \hat{\nu}$ then $L_{AC}(\psi) = L_{MP}(\psi)$, and worth noting that the NP adustment takes the simple form $r_\psi^{-1} \log[\{ j_{\nu\nu}(\hat{\psi}, \hat{\nu}) / j_{\nu\nu}(\psi, \hat{\nu})\}^{1/2}]$ depending only on the likelihood function. The key to seeing this is the well-known result that when $\hat{\nu}_\psi \equiv \hat{\nu}$ there is the simplification of the required sample-space derivative as $\partial^2 l(\psi, \hat{\nu}_\psi) / \partial\nu\,\partial\hat{\nu} = \partial^2 l(\psi, \hat{\nu}) / (\partial\nu)^2$; see for example Eqn. (5.14) of Barndorff-Nielsen & Cox (1994).

We now turn to two examples. In Ex. 2 there is no nuisance parameter, so the aim is not to illustrate our main result, but rather to provide some indication in the most fundamental setting of the adequacy of higher-order likelihood asymptotics. Example 3 then illustrates our main result regarding the NP adjustment and modified profile likelihood.

Example 2.

We consider independent observations $y_i \sim N(\theta,1)$ and two different stopping rules. The first is an example used by Woodroofe (1992), stopping for $n \le N/2$ at the first observation where $|r_{\theta=0}| > \eta n^{-1/2}$ and for $N/2 < n \le N$ where $|r_{\theta=0}| > 2\eta(N-n)^{1/2}/N$. The change in character at $N/2$ provides for the stopping boundary to close smoothly by the maximum number of trials N . The asymptotic index can be taken as η, with N increasing proportionately. We consider only one value of the asymptotic index for each part of this example, turning to formal asymptotics in Ex. 3. The notation $r_{\theta=0}$ in describing the stopping rule is to distinguish from the use of r_θ for testing other values of θ. Of course for this example $r_\theta^* = r_\theta$ for reference set when n is fixed, and so the adjustment $r_\theta^* - r_\theta$ reflects only the effect of the stopping rule. We raise this specific stopping rule mainly to compare r_θ^* to a related quantity proposed by Woodroofe (1992), of form $z_\theta^* = (r_\theta - n^{-1/2}\tilde{b})/(1+n^{-1}\tilde{c})$ with coefficients $\tilde{b},\tilde{c}$ being the asymptotic mean and standard deviation of r_θ, evaluated at the parameter estimate. He shows that under rather general conditions this statistic is in a ‘very weak sense’ standard normal to $o(n^{-1})$. It follows from considerations at the end of Appendix 2 that, provided our higher-order asymptotics are valid, z_θ^* and r_θ^* must agree to second order. As indicated by Woodroofe, under the general conditions for which he establishes the very weak convergence, there will be settings where z_θ^* is not to second order standard normal in the usual sense. Then neither Anscombe’s Theorem nor the theory underlying r_θ^* will apply. More practically, one might expect his statistic to perform somewhat better than r_θ^*, since it is specifically tailored to the sequential setting.

Substantial further development would be required for use of z^*_θ in the presence of nuisance parameters.

For this stopping rule, with $N = 72$ and $\eta = 9$, we obtain the results of Table 1, where the values for z^*_θ are taken from Woodroofe's paper. Those for r_θ and r^*_θ were obtained by simulation, using 10,000 datasets and 5,000 Monte Carlo trials for each of these to compute the covariance required by the Skovgaard approximation to r^*_θ. The conditions for Anscombe's Theorem hold for this example, and these results support our theoretical argument that for such an example r^*_θ is standard normal to second order, and to this order agrees with z^*_θ. For this example, and very specially, the symmetry of the model and stopping rule means that both $r^*_{\theta=0}$ and $z^*_{\theta=0}$ agree to second order with $r_{\theta=0}$. Thus $r_{\theta=0}$ is standard normal to second order, and the substantial improvement seen on this is due to adjustment of $O(n^{-1})$ not considered by our general theory.

Table 1. Part 1 of Ex. 2: Tail probabilities for various statistics.

True θ	Statistic	$pr < -1.96$	$pr < -1.645$	$pr > 1.645$	$pr > 1.96$
0.0	r_θ	0.051	0.101	0.101	0.050
	r^*_θ	0.024	0.051	0.050	0.024
	z^*_θ	0.025	0.051	0.048	0.024
0.5	r_θ	0.013	0.030	0.070	0.037
	r^*_θ	0.021	0.044	0.049	0.026
	z^*_θ	0.022	0.051	0.047	0.024
1.0	r_θ	0.015	0.032	0.060	0.031
	r^*_θ	0.023	0.051	0.051	0.023
	z^*_θ	0.026	0.054	0.053	0.021

For the remainder of this Ex. 2, we consider stopping rules seeming to us more practically important, having the general character of what is often used in clinical trials where the stopping boundary is not smoothly closed as above. There the intention is to stop usually at some maximum number of trials $n = N$, but earlier when the evidence against the hypothesis is substantial. We have in mind plans allowing at most *N* trials, stopping earlier if $| r_{\theta=0} | \ge \eta c_n$, with c_n decreasing modestly on the range $n = 1, \cdots, N$, and where in asymptotics η and N increase together. For Gaussian data as above, we consider only $N = 60$ and $\eta = 2$, with c_n decreasing linearly from $c_1 = 3/2$ to $c_N = 1$. We can avoid focus on testing only certain values of θ by considering confidence intervals. Figure 2 shows 90% equi-tailed confidence intervals for θ, based on various datasets on the stopping boundary, computed from the exact distributions of r_θ, and from standard normal approximations to r_θ and r_θ^*. The exact and r_θ^* upper limits are there indistinguishable, and for lower limits the upper curve is based on r_θ and the lower curve is exact. In terms of the *P*-values involved, those based on r_θ are too large by a factor of about two for $n = 10, 20$. The final 3 points for each limit are for $n = N = 60$, and r_θ-values of 2, 1, and 0.

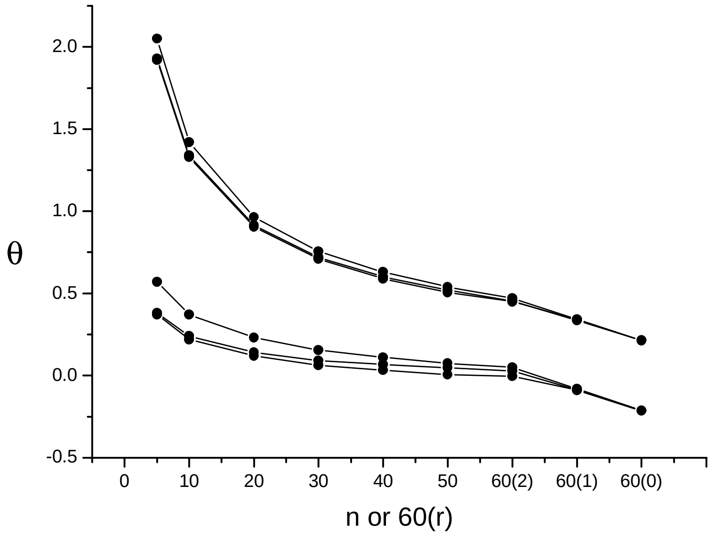


Figure 2. Confidence limits for data on the non-closed boundary: exact, first-order, and second-order.

Example 3.

This example illustrates for a sequential setting our main result regarding the *NP* adjustment, and also some of the more general asymptotic aspects. We consider again inference about the shape parameter in the Weibull model, but now with identically distributed observations involving only scale and shape parameters, and no censoring. Both the stopping rule and the inference pertain to testing that the shape parameter is unity. The stopping boundary is that of the first part of Ex. 2, with $\eta = N/8$, but using for that purpose the score test rather than the likelihood ratio statistic. Since employing a first-order test in the stopping rule is inappropriate for very small sample sizes, we do not stop when $n < 4$. The higher-order adjustments are computed by simulation as indicated in Ex. 1. For perspective, both the *NP* and *INF* adjustments are moderately important even for $N = 100$, where the probability of stopping prior to $n = N$ is 19% and $E(n) = 85$. In that case the *NP* adjustment takes values around -0.1 when $|r_{\psi=1}|$ is in the range $1.5 - 2.0$, the *INF* adjustment is around -0.2 when r_ψ is in the range $1.5 - 2.0$, and is around $+0.1$ when $-r_\psi$ is in that range.

Our main result is confirmed by finding that the difference in *NP* adjustments for the sequential setting and when treating *n* as fixed at the stopping value is around $\pm 10^{-3}$ for all datasets. Plots show this having the character of being second order in $E(n)$, but are not shown since the differences are so small.

Figure 3 shows as a function of $E(n)$ a summary measure of departure from standard normality of r_ψ, r_ψ^* when *n* is considered as fixed at the stopping value, and r_ψ^* for the sequential setting. The curves from top to bottom are for the three statistics in the order just stated. The summary measure is the average of absolute relative error in tail probabilities at 0.025, 0.05, 0.10, and 0.15 in each direction. Also shown, to investigate second-order convergence, is the summary error measure multiplied by $E(n)$. The convergence of r_ψ^* correctly computed for the sequential setting is, as theory suggests, of second order. Although r_ψ and the incorrect fixed-*n* r_ψ^* perform poorly, the rate of improvement at the largest values of $E(n)$ appears better than the first-order that is expected. Generally, we find the nature of the stopping boundary to have rather complicated effects on concrete details of asymptotic behaviour, a point to which we will return in the Discussion.

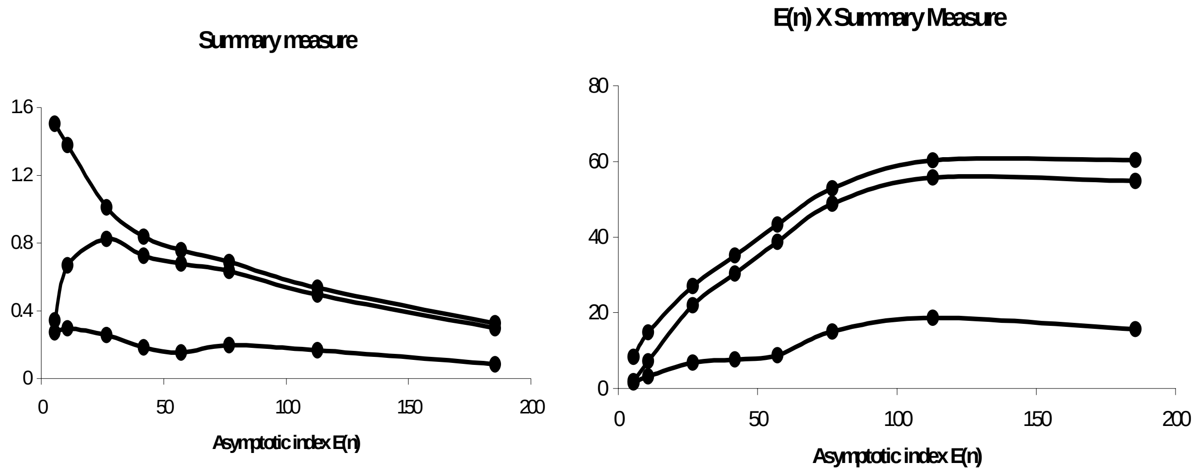


Figure 3. Comparison to standard normal of the distributions of r_ψ, r^*_ψ, and of r^*_ψ when *n* is considered as fixed at the stopping value.

Figure 4 shows quartiles of the difference in *INF* adjustments computed for the sequential setting and when *n* is treated as fixed at the stopping values. In contrast to what occurs with the *NP* adjustment, this difference reflects the effect of the stopping rule and is of first order in $E(n)$. A further plot not shown indicates that the difference is not also of second order.

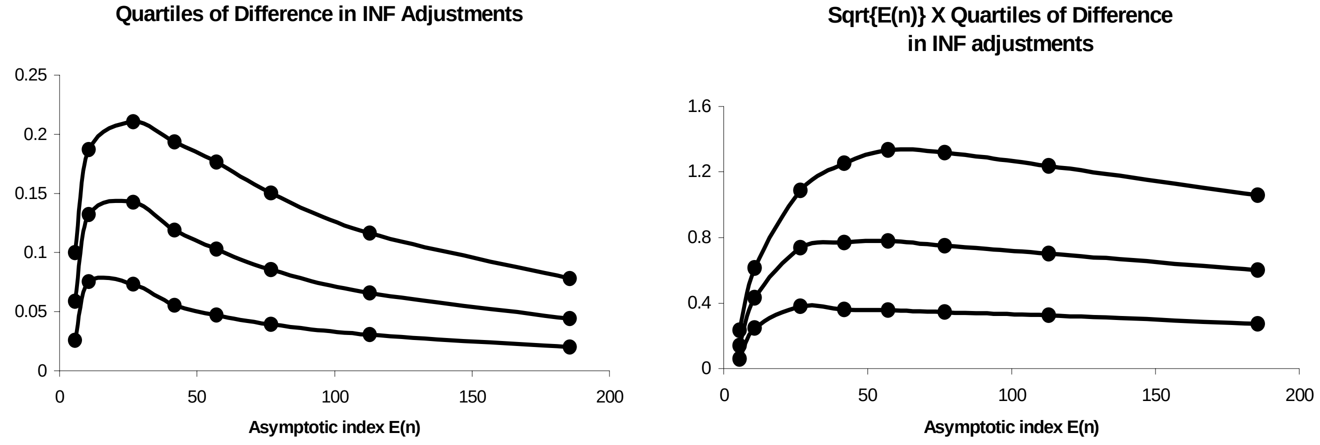


Figure 4. Quartiles of distributions of differences in *INF* adjustments for the sequential reference set and when *n* is considered as fixed.

5. DISCUSSION

For the two settings considered in this paper, the *NP* adjustment and modified profile likelihood are to second order independent of the reference set. The question arises of whether the *NP* adjustment ever depends on the reference set, and we suspect that there are instances where it does. In both settings of this paper the matter hinges on likelihood-related martingales whose contributions are the same for the reference sets considered, and it seems likely that generalizations would require this. However, as further indicated below, even from a martingale view the arguments for censored data and for stopping rules seem fundamentally different, and it is unlikely that there is a single general argument.

Modified profile likelihood is an attractive concept that needs more attention in practice. The work of Brazzale (1999), who implemented S-plus routines for certain important applications, should help with this. Practical implementation can be carried out substantially more generally using the method indicated in § 9.5.4 of Severini (2000), which basically corresponds to the *NP* part of the Skovgaard approximation used in this paper.

From a frequency viewpoint a primary attraction of the modified profile likelihood is that the *INF* adjustment is usually small in practice, and use of the modified profile likelihood is then tantamount to use of r_{ψ}^{*}. Ordinary asymptotics is rather misleading when the *NP* adjustment is large and the *INF* adjustment is small, and more relevant asymptotics involve allowing the number of parameters to increase along with the sample size; see for example Sartori (2003) and references there.

Modified profile likelihood is attractive on Bayesian grounds, where it can circumvent need for a prior distribution on nuisance parameters. The argument of Sweeting (1987) referred to in Appendix I is important in this, but more work in that direction is needed. His condition that the nuisance parameter can be taken as both orthogonal to and *a priori* independent of ψ is quite restrictive, and moreover his result does not, even then, distinguish between use of the Cox-Reid approximate conditional likelihood and the Barndorff-Nielsen modified profile likelihood. Since the latter is invariant to the representation of the nuisance parameter, Bayesian arguments leading more directly to it should involve invariance of prior distributions.

In regard to sequential experiments, it is clear to us that the formulation of the stopping rule and inferential matters should be considered more separately than they usually are. In particular, we believe it is neither necessary nor desirable to identify some portion of the stopping boundary as a rejection region. Others have come to essentially this view (Armitage, 1957; Whitehead, 1999), but it seems fair to say that much work, both classical and recent, on sequential experiments is confusing in these respects.

The first-order convergence implied by Anscombe's Theorem is often quite slow. Although second-order methods usually improve substantially on this, they seem unlikely to provide practical methods because they are difficult to compute and direct simulation of the distribution of r_ψ is much simpler. Moreover, the precise performance of r_ψ^* depends in complicated ways on specific aspects of the stopping boundary. Our view is that a primary value of higher-order asymptotics for sequential settings, and indeed more generally, is for theoretical understanding of the structure of inference. That is, the effects of fitting nuisance parameters, the role of choice of reference set, connections between frequency and likelihood-Bayesian inference, and so forth. From this view it is not critical that one may not want to actually compute r_ψ^*, or rely strongly on it, for sequential settings.

For censored data, that higher-order inference is independent of the censoring model does not in itself resolve how to make second-order adjustments in practice. For the types of censoring considered in § 3, our argument does suggest a method to use, which essentially involves the approximation to r_ψ^* suggested by Severini (1999) described briefly in § 3. However, we are not certain that this is the best method, and further work would be useful.

It appears that our results for both the *NP* and *INF* adjustments extend to general censoring models, although there are some logical subtleties involved. Censoring models as considered in § 3 are essentially referred to as Type I, and those where they are taken as certain of the ordered failure times are referred to as Type II. It turns out to be important in second-order considerations that a given dataset will not be exactly compatible with both Type I and Type II censoring; the issue being whether censoring times coincide with failure times. Thus whether ideal inference from a given dataset is the same for Type I and II censoring models seems to be not quite the right question. Indeed, direct calculations show that for instances of the two types of models with censoring patterns the same in expectation, the values of $\mathrm{cov}(U, J)$ typically fail to agree to first order. However, the more fundamental question is whether knowledge of the censoring model is required for ideal inference from a given dataset, and the answer seems to be that it is not.

The issues can be dealt with in a general manner through martingale considerations, a useful general reference being Andersen, et al. (1993). Although we have not fully investigated the matter, our analysis including numerical results for a simple example indicates that there is a first-order approximation to $\mathrm{cov}(U, J)$, namely the compensator of UJ, depending on the observed censoring but not on knowledge of the censoring model. This compensator is sensitive to whether or not the censoring times occur exactly at failure times, and so the estimator can be said to adapt to whether censoring is of Type I or II. We should add that it is not always true in martingale considerations that such a compensator involves only the observed data and not the model for it; see for example Barndorff-Nielsen & Sørensen (1994) regarding what they refer to as the 'incremental expected information'. That the compensator of UJ for our setting depends only on the observed data, and indeed its very existence, is special to counting processes arising for survival data. This is why the argument just alluded to fails to apply for the considerations in §4.

Returning to general considerations, we have not intended in this paper to take a position on the merits of restricting to first-order methods that do not depend on the reference set; for example relying on Anscombe's Theorem for sequential experiments. Rather, we simply aim to clarify what is involved in such issues, by quantification to an extent reference set effects. However, our results regarding modified profile likelihood bear on such matters, since it is easier to recommend general use of this after finding that it is widely independent of the reference set.

Appendix 1

We take it for granted that ideal inference should fundamentally involve likelihood ratios, but in the presence of nuisance parameters there are various pseudo-likelihood functions for the interest parameter that should be considered. It is indicated here that P-values computed from the exact or second-order distribution of

$r_\psi(y)$ agree to second order with those computed similarly, but where $r_\psi(y)$ is replaced by the generalized likelihood ratio statistic computed from the modified profile likelihood or from a Bayesian likelihood L_B obtained by integrating out the nuisance parameter according to any smooth prior distribution. The point is that the various statistics provide to second order the same ordering of *datasets* according to evidence against the hypothesis, even though, as is well known, this degree of equivalence does not obtain when comparing the pseudo-likelihoods as functions of ψ. For example, as functions of ψ, L_{MP} and L_P generally differ by first order. Comprehensive comparisons of the pseudo-likelihoods as functions of ψ are given by Severini (1998). It is further indicated at the end of Appendix 2 that to second order $r_\psi(y)$ and $r_\psi^*(y)$ provide the same ordering of datasets.

Specifically, we will show that for $L_{MP}(\psi)$ the logarithmic directional likelihood ratio is to second order of form $r_\psi - n^{-1/2}b(\theta)$, so the ordering of datasets is the same as when using r_ψ, the directional likelihood ratio from $L_P(\psi)$. It is the function of $b(\theta)$ which results in the usually considered first-order difference between modified profile and profile likelihoods as functions of ψ. We then show that this also holds for integrated likelihoods $L_B(\psi)$ with the function $b(\theta)$ depending also on the prior distribution, leading to the same overall conclusion.

Indeed, a strong case can be made, in view of the Wald complete class theory, that ideal frequency inference should be based on ordering of datasets by ratios of an integrated likelihood

$$L_B(\psi) = \int L(\psi,\nu)\pi(\nu \mid \psi)\, d\nu$$

for some prior distribution $\pi(\psi,\nu)$. We will show that to second order such an ordering does not depend on the prior, provided it is suitably smooth.

Considering the Cox & Reid (1987) proposal given in Appendix II, and combining use of a Laplace approximation given by Sweeting (1987) with the second-order equivalence noted by Barndorff-Nielsen (1987) of L_{MP} and the Cox-Reid proposal, we can express to that order

$$L_B(\psi) \doteq \pi(\hat{\nu}_\psi \mid \psi)\, L_{AC}(\psi) \doteq \pi(\hat{\nu}_\psi \mid \psi)\, L_{MP}(\psi) = \pi(\hat{\nu}_\psi \mid \psi)\, M(\psi) L_P(\psi)$$

where $M(\psi)$ is defined in §2.

First we consider the modified profile likelihood $L_{MP}(\psi) = M(\psi) L_P(\psi)$. Following Barndorff-Nielsen (1986), and discussed further at the end of Appendix 2, we can express via a function $Q = O(1)$ the modified profile likelihood adjustment as

$$\log\{M(\hat{\psi})/M(\psi)\} = n^{-1/2} Q(\psi, \hat{\nu}_\psi, a)\, r_\psi + O(n^{-1}) = n^{-1/2} \bar{Q}(\theta)\, r_\psi + O(n^{-1})\,.$$

The replacement of $Q(\psi, \hat{\nu}_\psi, a)$ by the function $\bar{Q}(\theta)$ not depending on the data is crucial to the argument here. It is immaterial throughout arguments here whether pseudo-likelihood ratios use $\hat{\psi}$, the $\hat{\psi}_M$ maximizing L_{MP}, or the $\hat{\psi}_B$, since all these agree to $O(n^{-1})$ without standardization, and lead to distinctions of second order in likelihood ratios. The above expansion means that to second order $\log\{L_{MP}(\psi)/L_{MP}(\hat{\psi}_M)\}$ is a quadratic function of r_ψ with coefficients depending on θ, and completing the square gives the result claimed above for the case of L_{MP}.

Turning to the integrated likelihood, note first that both L_{MP} and L_B are invariant to the representation of ν, so we may with no loss take that as orthogonal to ψ. As noted by Sweeting (1987), if $\pi(\nu \mid \psi)$ does not depend on ψ then the contribution to the likelihood ratio resulting from that term is unity to second order since $\hat{\nu}_\psi - \hat{\nu} = O(n^{-1})$, and our final result for L_B is the same as for L_{MP}. But that *a priori* independence, along with orthogonality, is a very strong assumption. When $\pi(\nu \mid \psi)$ depends on ψ, then L_B and L_{MP} differ by first order as functions of ψ, but we now show that our claim regarding ordering of datasets still holds.

The point is how the factor $\pi(\hat{\nu} \mid \hat{\psi})/\pi(\hat{\nu}_\psi \mid \psi)$ modifies $L_{MP}(\hat{\psi})/L_{MP}(\psi)$. Since $\hat{\nu}_\psi - \hat{\nu} = O(n^{-1})$, writing $g(\psi,\nu) = \partial \log \pi(\nu \mid \psi)/\partial \psi$ we have to second order that $\pi(\hat{\nu} \mid \hat{\psi})/\pi(\hat{\nu}_\psi \mid \psi) = 1 + g(\psi, \hat{\nu})(\hat{\psi} - \psi) = 1 + g(\psi,\nu)(\hat{\psi} - \psi) = \exp\{g(\psi,\nu)(\hat{\psi} - \psi)\}$. It follows routinely from approximating the log profile likelihood as quadratic in ψ that $(\hat{\psi} - \psi) = i_{\psi|\nu}^{-1/2}(\theta)\ r_\psi + O(n^{-1})$, where $i_{\psi|\nu}(\theta)$ is the expected adjusted information. Combining these results with those for $L_{MP}(\psi)$ means that $l_B(\hat{\psi}) - l_B(\psi)$ is a quadratic function of r_ψ with coefficients depending on θ and now, through g, on the prior as well. Completing the square shows that to second order the logarithmic directional likelihood ratio is of form $r_\psi - n^{-1/2} b(\theta, g)$, providing the claimed result for L_B.

Appendix 2

Everything to follow can be obtained by combining results in Sects. 6.6 and 8.2 of Barndorff-Nielsen & Cox (1994) with those in Sects.7.4.1, 7.5.4 of Severini (2000). However, it is useful to gather together what is needed for this paper.

The material of the next few paragraphs was developed by Barndorff-Nielsen (1986), but it is clearer to refer to the argument as presented in Severini (2000). The derivation of r_ψ^* involves the likelihood ratio approximation to the density $p(\hat{\theta} \mid a; \theta)$, a simple derivation of which is indicated in Severini's §6.3.2 and is considered further in Appendix III of this paper. From this, one can transform to a second-order approximation to the distribution of r_ψ as indicated in Severini's §7.4.1, arriving at $pr\left\{r_\psi(y) \le r_\psi(y_{obs}) \mid a; \theta\right\} = \Phi\left\{r_\psi^*(y_{obs})\right\}$, where Φ is the standard normal cumulative distribution function and r_ψ^* is as given below. Barndorff-Nielsen (1986) showed that r_ψ^* is stochastically independent of a to second order, and it follows from this and considerations at the end of this Appendix to this order r_ψ is also stochastically independent of a, so that unconditionally as well ,
$pr\left\{r_\psi(y) \le r_\psi(y_{obs}); \theta\right\} = \Phi\left\{r_\psi^*(y_{obs})\right\}$.

Following notation of Barndorff-Nielsen & Cox, and noting that $M(\psi)$ of §2 is the same as C_ψ^{-1} below, we can write eqn. (4) of the text more explicitly as

$$r_\psi^* = r_\psi + r_\psi^{-1} \log(C_\psi^{-1}) + r_\psi^{-1} \log(\tilde{u}_\psi / r_\psi) \qquad \text{(A2.1)}$$

where

$$\tilde{u}_\psi = \hat{j}_{\psi|\nu}^{-1/2}\, \partial / \partial \hat{\psi}\, \{l_P(\psi) - l_P(\hat{\psi})\}$$
$$C_\psi^{-1} = \mid \partial^2\, l(\psi, \hat{\nu}_\psi) / \partial \nu \partial \hat{\nu} \mid / \left\{\mid j_{\nu\nu}(\psi, \hat{\nu}_\psi) \mid \mid \hat{j}_{\nu\nu} \mid\right\}^{1/2}$$

Here $\hat{j}$ denotes the observed information at the maximum likelihood estimator, $\hat{j}_{\psi|\nu}$ is the adjusted information there. For partial differentiation the data specifying the likelihood, suppressed in notation above, are represented as $(\hat{\psi}, \hat{\nu}_\psi, a)$. The two adjustment terms in (A2.1) are each $O(n^{-1/2})$ and are invariant to representation of the nuisance parameter. In related theory Barndorff-Nielsen has defined a modified profile likelihood

$$\begin{aligned} L_{MP}(\psi) &= |\partial\hat{\nu}_{\psi}/\partial\hat{\nu}|^{-1} |\, j_{\nu\nu}(\psi,\hat{\nu}_{\psi})|^{-1/2}\ L_P(\psi) \\ &\propto C_{\psi}^{-1}\, L_P(\psi) \end{aligned} \qquad \text{(A2.2)}$$

The final step there follows from Eqn. (8.21) of Barndorff-Nielsen & Cox (1994). As noted in § 2, these considerations lead to designation of the two adjustment terms in (A2.1) as the nuisance parameter and information adjustments. It is the term $|\partial\hat{\nu}_{\psi}/\partial\hat{\nu}|$ that renders this invariant to the representation of the nuisance parameter, but this is difficult to compute. If the nuisance parameter is chosen to be orthogonal to ψ, then this term is $1+O(n^{-1})$, and omitting it in that case leads to the Cox & Reid (1987) approximate conditional likelihood

$$L_{AC}(\psi) = |\, j_{\nu\nu}(\psi,\hat{\nu}_{\psi})|^{-1/2}\ L_P(\psi)\,. \qquad \text{(A2.3)}$$

The problem with this is that if ν is orthogonal to ψ, then so is any choice $\lambda = g(\nu)$, introducing a great deal of arbitrariness in L_{AC} even though all forms agree to second order; see, for example, Severini (2000, Example 9.13).

The Skovgaard approximations to the sample space derivatives required for (A2.1) are as follows. Writing $U(\theta)$ for the score statistic evaluated at a parameter value θ, define

$$\begin{aligned} \widehat{\text{cov}}(\hat{U},\tilde{U}) &= \text{cov}_{\theta_1}\{U(\theta_1),U(\theta_2)\} \\ \widehat{\text{cov}}(\hat{U},\widetilde{\Delta l}) &= \text{cov}_{\theta_1}\{U(\theta_1),l(\theta_2)-l(\theta_1)\} \end{aligned}$$

where, following computation of the expectations, θ_1 and θ_2 are respectively evaluated at the unconstrained and constrained maximum likelihood estimators $\hat{\theta}$ and $\tilde{\theta}_{\psi} = (\psi,\hat{\nu}_{\psi})$. Writing $\hat{j}$ and $\hat{i}$ for the observed and expected information evaluated at $\hat{\theta}$, the Skovgaard approximations for quantities involved in A2.1 are

$$\begin{aligned} \partial^2 l(\psi,\hat{\nu}_{\psi})/\partial\theta\partial\hat{\theta} &= \widehat{\text{cov}}(\hat{U},\tilde{U})\,\hat{i}^{-1}\hat{j} \\ \partial/\partial\hat{\theta}\,\{l(\psi,\hat{\nu}_{\psi})-l(\hat{\psi},\hat{\nu})\} &= \widehat{\text{cov}}(\hat{U},\widetilde{\Delta l})\,\hat{i}^{-1}\hat{j} \end{aligned}$$

In § 3 we indicated how each of these can be expressed as the sum of a term depending only on the likelihood function and term of one smaller order depending on the reference set. Going from the second expression above to the quantity $\tilde{u}_\psi$ referred to following (A2.1) is not totally straightforward because the quantity $\hat{\nu}_\psi$ is held fixed in the partial differentiation; formulas for this are given in (6.106–107) of Barndorff-Nielsen & Cox (1994).

It is useful for many purposes to note that, as shown by Barndorff-Nielsen (1986), r_ψ^* can to third order be expressed as $r_\psi^* = \{r_\psi - n^{-1/2} b(\psi, \hat{\nu}_\psi, a)\} / \{1 + n^{-1} c(\psi, \hat{\nu}_\psi, a)\}$. Moreover, to second but not third order the arguments $(\psi, \hat{\nu}_\psi, a)$ there can be replaced by θ, and of course to that order the denominator is not relevant. Consequently, the orderings of datasets according to evidence against the hypothesis ψ given by $r_\psi(y)$ and $r_\psi^*(y)$ are to second but not third order equivalent. More fundamentally, in regard to the relation $r_\psi^* = r_\psi + r_\psi^{-1} \log(u_\psi / r_\psi)$, Barndorff-Nielsen considered the expansion $\log(u_\psi / r_\psi) = r_\psi n^{-1/2} Q_1(\psi, \hat{\nu}_\psi, a) + r_\psi^2 n^{-1} Q_2(\psi, \hat{\nu}_\psi, a)$, noting that the coefficients of the above affine transformation are then $b = Q_1$ and $c = Q_2 / 4 - Q_1^2 / 2$. He then argued that the functions Q_1 and Q_2 can be expanded in powers of $\hat{\nu}_\psi - \nu$ and $a - E(a)$ with leading terms $\bar{Q}_1(\theta)$ and $\bar{Q}_2(\theta)$, leading to the expression for r_ψ^* stated above. When $\log(u_\psi / r_\psi)$ is decomposed as in (A2.1), analogous arguments for replacing $(\psi, \hat{\nu}_\psi, a)$ by θ apply to each term of the decomposition.

Appendix 3

We indicate somewhat heuristically here why the considerations of § 2 apply to sequential settings, provided that the stopping rule is such that ordinary likelihood-based methods are valid to second order. Some idealizations will ordinarily be required so that densities to follow can be taken with respect to Lebesgue measure. We begin with one-parameter models and then indicate what is required with nuisance parameters. The argument to follow is not restricted to sequential settings, and clarifies the role of conditioning on ancillaries for inference based on r_ψ^*.

Likelihood asymptotics begin with the identity

$$\begin{aligned} p(\hat{\theta}\,|\,a;\theta) &= \frac{p(\hat{\theta}\,|\,a;\theta)}{p(\hat{\theta}\,|\,a;\hat{\theta})}\, p(\hat{\theta}\,|\,a;\hat{\theta}) \\ &= \exp(-r_\theta^2/2)\, p(\hat{\theta}\,|\,a;\hat{\theta}) \end{aligned}$$

with the second form holding when a is exactly ancillary. When a is ancillary to first order then the result holds to second order for moderate deviations of θ. The usual approach is to assume that $p(\hat{\theta}\,|\,a;\theta)$ is normal to first order, and then to replace $p(\hat{\theta}\,|\,a;\hat{\theta})$ by its value $(2\pi j)^{-1/2}$ from a second-order Edgeworth expansion, see for example, § 6.3.2 of Severini (2000). We will avoid that step here, assuming for the moment only that there is a first-order ancillary so that that above identity is applicable.

Although conditions on $p(\hat{\theta}\,|\,a;\hat{\theta})$ are essential for the likelihood ratio approximation to the distribution of $\hat{\theta}$, they are not for approximating the distribution of r_θ. Transforming the above relation to the density for r_θ leads to the relations

$$\begin{aligned} p(r_\theta\,|\,a;\theta) &= \left[\,|\,\partial\hat{\theta}/\partial r_\theta\,|\;\hat{j}^{1/2} e^{-r_\theta{}^2/2}\right]\left[\,\hat{j}^{-1/2} p(\hat{\theta}\,|\,a;\hat{\theta})\right] \\ &= \left[\,r_\theta/u_\theta\;\; e^{-r_\theta{}^2/2}\right]\left[\,\hat{j}^{-1/2} p(\hat{\theta}\,|\,a;\hat{\theta})\right] \\ &= \left[\,r_\theta/u_\theta\;\; e^{-r_\theta{}^2/2}\right] p(r_{\hat{\theta}}\,|\,a;\hat{\theta}) \end{aligned}$$

where u_θ is as defined in Appendix 2, with the profile loglikelihood l_P replaced by the loglikelihood, and using the fact that $d\,r_\theta/d\,\hat{\theta}$ evaluated at $\theta=\hat{\theta}$ is $\hat{j}^{-1/2}$. The final expression on the last line means the density $p(r_\theta\,|\,a;\theta)$ when θ is taken as $\hat{\theta}$, the argument there is zero. As noted at the end of Appendix 2 there is a function $b(\theta)$ such that $r_\theta^* = r_\theta + r_\theta^{-1}\log(u_\theta/r_\theta)$ is to second order equal to $r_\theta - n^{-1/2}b(\theta)$. Thus we can integrate both sides of the above expression with respect to the distribution of a to obtain

$$p(r_\theta\,;\theta) = \left[\,r_\theta/u_\theta\;\; e^{-r_\theta{}^2/2}\right] p(r_{\hat{\theta}}\,;\hat{\theta})\,. \qquad \text{(A3.1)}$$

Then when the stopping rule is such that r_θ is to first order standard normal, it is a minor further assumption that its distribution has a second-order Edgeworth expansion, according to which $p(r_{\hat{\theta}}\,;\hat{\theta}) = (2\pi)^{-1/2}$ since the skewness term vanishes at

$r_{\hat{\theta}} = 0$. With this substitution in (A3.1), approximation (3) of the main text follows rather directly from completing the square in the logarithm of the term in brackets and then integrating.

When there is a nuisance parameter, the distribution $p(r_\theta \mid a;\theta)$ above is replaced by $p(r_\psi, \hat{v}_\psi \mid a;\theta)$, which involves another Jacobian in replacing $\hat{v}$ by $\hat{v}_\psi$. Then, proceeding as in § 7.4.1 of Severini (2000), one integrates out $\hat{v}_\psi$. The result is then as in (A3.1), where now u_ψ is as defined in Appendix 2 using the profile likelihood.

The second-order approximation to the marginal distribution of r_ψ obtained is that given by (3) of the main text, without needing to assume as in the usual development that $p(\hat{\theta} \mid a;\theta)$ is normal to first order, but only that $p(r_\psi;\theta)$ has this property.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York, Springer-Verlag.

Anscombe, F. J. (1952). Large sample theory of sequential estimation. *Proc. Camb. Phil. Soc.* **48**, 600-7.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9-26.

Armitage, P. (1991). Sequential Methods, Ch. 6 in *Statistical Theory and Modelling, In Honour of Sir David Cox, FRS.* Eds Hinkley, D. V., Reid, N. & Snell, E. J. London, Chapman & Hall.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-65.

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized likelihood ratio. *Biometrika* **73**, 307-22.

Barndorff-Nielsen, O. E. (1987). Contribution to Discussion of Cox & Reid, Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B.* **49**, 1-39.

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

Barndorff-Nielsen, O. E. & Cox, D. R. (1984). The effect of sampling rules on likelihood statistics. *Int. Statist. Rev.* **52**, 309-26.

Barndorff-Nielsen, O. E. & Sørensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes. *Int. Statist. Rev.* **62**, 133-65.

Brazzale, A. R. (1999). Approximate conditional inference in logistic and loglinear models. *J. Comput. and Graph. Statist.* **8**, 653-61.

Coad, D. S. & Woodroofe, M. B. (1996). Corrected confidence intervals after sequential testing with applications to survival analysis. *Biometrika* **83**, 763-78.

Cook. T. D. (2002). *P*-value adjustment in sequential clinical trials. *Biometrics* **58**, 1005-11.

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

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

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**, 67-76.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457-87.

Fraser, D. A. S., Reid, N. & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249-64.

Fraser, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-39.

Grambsch, P. (1983). Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Ann. Statist.* **11**, 68-77.

Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference.* Singapore, World Scientific.

Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc.B* **54**, 701-37.

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

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd Ed. New York, John Wiley & Sons.

Reid, N. (1996). Likelihood and higher-order approximations to tail areas: A review and annotated bibliography. *Canad. J. Statist.* **24**, 141-66.

Reid, N. (2004). Asymptotics and the theory of inference, to appear in *Annals of Statistics.*

Rosner, G. L. & Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75**, 723-9.

Sartori, N., Bellio, R., Salvan, A. & Pace, L. (1999). The directed modified profile likelihood in models with many parameters. *Biometrika* **86**, 735-42.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533-49.

Severini, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85**, 507-22.

Severini, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika* **86**, 235-47.

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

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika* **65**, 341-9.

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

Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145-65.

Sweeting, T. J. (1987). Contribution to Discussion of Cox & Reid, Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B.* **49**, 1-39.

Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88**, 657-75..

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statist. in Med.* **18**, 2271-86.

Whitehead, J., Todd, S. & Hall, W. J. (2000). Confidence intervals for secondary parameters following a sequential test. *J. Roy. Statist. Soc. B.* **62**, 731-45.

Woodroofe, M. (1992). Estimation after sequential testing: A simple approach for a truncated sequential probability ratio test. *Biometrika* **79**, 347-53.

Woodroofe, M. & Keener, R. (1987). Asymptotic expansions in boundary crossing problems. *Ann. Prob.* **15**, 102-14.