Bernoulli **7**(3), 2001, 473–485

Likelihood computations without Bartlett identities

PER ASLAK MYKLAND

Department of Statistics, University of Chicago, Chicago IL 60637, USA.
E-mail: mykland@galton.uchicago.edu

The signed square root statistic R is given by $\mathrm{sgn}(\hat\theta-\theta)(l(\hat\theta)-l(\theta))^{1/2}$, where l is the log-likelihood and $\hat\theta$ is the maximum likelihood estimator. The pth cumulant of R is typically of the form $n^{-p/2}k_p+O(n^{-(p+2)/2})$, where n is the number of observations. This paper shows how to symbolically compute k_p without invoking the Bartlett identities. As an application, we show how the family of alternatives influences the coverage accuracy of R.

Keywords: Bartlett correction; convergence of cumulants; unconditional accuracy

1. Introduction

A central object in likelihood theory is the so-called R statistic, otherwise known as the signed square root of the likelihood ratio statistic, $R=\mathrm{sgn}(\hat\theta-\theta)(l(\hat\theta)-l(\theta))^{1/2}$, where l is the log-likelihood and $\hat\theta$ is the maximum likelihood estimator. For discussions of this statistic, see, for example, Barndorff-Nielsen (1986; 1991) and McCullagh (1984). To assess and improve the accuracy of R, substantial effort has gone into computing its cumulants, usually by a method of combining stochastic Taylor expansions with Bartlett identities; see, in particular, Bartlett (1953a; 1953b), Lawley (1956), Shenton and Bowman (1977, Chapter 3), McCullagh (1984; 1987, p. 202), Peers and Iqbal (1985), Skovgaard (1986), McCullagh and Tibshirani (1990), DiCiccio and Romano (1989), DiCiccio *et al*. (1991), DiCiccio and Stern (1993; 1994) and Mykland (1994; 1995a; 1995b).

Calculations with Bartlett identities are, however, notoriously cumbersome. One solution to this problem lies in computer algebra; see Stafford and Andrews (1993), Andrews and Stafford (1993; 2000), Stafford (1994; 1995), and Stafford *et al*. (1994). In this paper, we shall suggest a different approach to deal with this issue.

The purpose of the following is to present a way of deriving the cumulants of R in a computationally much simpler manner, bypassing the traditional machinery. Specifically, if R_n is the R statistic based on n observations, and if $\mathrm{cum}_p(R_n)$ is the pth cumulant of R_n (see, for example, McCullagh 1987, Chapter 2), we shall see below that

$$\mathrm{cum}_p(R_n)=\delta_{2,p}+n^{-p/2}k_p+O(n^{-(p+2)/2}). \tag{1.1}$$

Here $\delta_{2,p}=1$ for $p=2$ and $\delta_{2,p}=0$ otherwise. We shall provide a formula for the generating function of the k_ps with the help of (2.1), (2.5) and Theorem 2 below. We then show how to implement the procedure by symbolic computation (Section 2). As an

1350–7265

application, we discuss (Section 3) how these results can be used to analyse the effect of the alternative on the null distribution of R, and how this affects the difference between nominal and actual coverage of confidence intervals.

To establish (1.1), note that, subject to regularity conditions, an asymptotically standard normal statistic T_n satisfies $\text{cum}_p(T_n) = n^{-(p-2)/2}\kappa_p + n^{-p/2}k_p + O(n^{-(p+2)/2})$, with $\kappa_p = \delta_{2,p}$ for $p = 1, 2$. This follows from general results on the expansions of cumulants in Wallace (1958), Bhattacharya and Ghosh (1978) and Hall (1992). It now follows from Theorem 1 of Mykland (1999) that $\kappa_p = 0$ for $p \geqslant 3$ when $T_n = R_n$. The results in the latter paper are related to large-deviation results for R statistics in Barndorff-Nielsen and Wood (1998), Jensen (1992; 1995, Chapter 5; 1997) and Skovgaard (1990; 1996).

2. The main formula

2.1. Theoretical development

Suppose that a one-parameter family of probabilities P_β is given, and that $l_n(\beta)$ is the log-likelihood based on n observations. Let the k_ps from (1.1) be defined when the cumulants are taken under distribution P_{β_0}, and denote the exponential of their generating function by

$$\xi(h) = \exp\left\{ k_1 h + \frac{1}{2} k_2 h^2 + \frac{1}{3!} k_3 h^3 + \dots \right\}. \tag{2.1}$$

Our purpose in this section is to display a formula for ξ. Consider the density f_n of R_n, and the cumulant generating function K_n, both under P_{β_0}. The latter is given by

$$K_n(h) = h\text{E}(R_n) + \frac{1}{2} h^2 \text{var}(R_n) + \frac{1}{3!} h^3 \text{cum}_3(R_n) + \dots, \tag{2.2}$$

and, under regularity conditions on the remainder term in (1.1), one obtains

$$K_n(n^{1/2}h) = \tfrac{1}{2}nh^2 + \log \xi(h) + O(n^{-1}). \tag{2.3}$$

The saddlepoint approximation to f_n has the form

$$f_n(n^{1/2}h) = \frac{1}{(2\pi K_n''(\hat{\tau}_n))^{1/2}} \exp(K_n(\hat{\tau}_n) - \hat{\tau}_n K_n'(\hat{\tau}_n))(1 + o(1)), \tag{2.4}$$

where $K_n'(\hat{\tau}_n) = n^{1/2}$. For the sample mean, this goes back to Daniels (1954), and a version of (2.4) for general statistics is embodied in Theorem 1 of Chaganty and Sethuraman (1985). See also the development in Jensen (1995). Since we are dealing with the signed square root statistic R_n, whose cumulants are of the form (1.1), one can expand (2.4) in a Taylor series to obtain

$$\xi(h) = \lim_{n\to\infty} \frac{f_n(n^{1/2}h)}{\phi(n^{1/2}h)}, \tag{2.5}$$

ϕ being the standard normal density. Equation (2.5) is important because it can be combined with Theorem 1 below to find an explicit formula for ξ (Theorem 2).

Definition. *Let $l_n = l_n(\beta)$ be the log-likeliood based on n observations. $\dot{l}_n$, $\ddot{l}_n$ and $l_n^{(p)}$ are derivatives with respect to the (scalar) parameter β. Also set*

$$\tilde{l}(\beta) = \lim_{n\to\infty} (l_n(\beta) - l_n(\beta_0))/n \tag{2.6}$$

and

$$J(\beta) = -\lim_{n\to\infty} \ddot{l}_n(\beta)/n \tag{2.7}$$

where the limits are in probability under P_β. Finally, let

$$h(\beta) = \sqrt{2}\,\mathrm{sign}(\beta - \beta_0)\tilde{l}(\beta)^{1/2}. \tag{2.8}$$

Our first result approximates the density of R_n.

Theorem 1. *In a curved exponential family, and under the assumptions of Jensen (1997) and of Appendix 1,*

$$f_n(r) = \phi(r)J^{1/2}(\beta)\frac{\partial\beta}{\partial h}(h)\{1 + O(n^{-1/2})\} \tag{2.9}$$

in a large-deviation region $|h| \leqslant c$, with $h = r/\sqrt{n}$.

For the proof, see Appendix 1. The above now combines with (2.5) to yield the following result.

Theorem 2. *Subject to regularity conditions, ξ is given under P_{β_0} by*

$$\xi : h \to J^{1/2}\frac{\partial\beta}{\partial h}. \tag{2.10}$$

In the case of curved exponential families, the latter theorem is a direct corollary to the former. For smooth families that are not in the form of a finite curved family, one can proceed as follows. If R is the signed square root statistic in a sufficiently smooth likelihood family, one can write $R = \bar{R} + O_p(n^{-(p-1)/2})$, where $\bar{R}$ is the R statistic in a $(p+2,\ 1)$ curved exponential family. (see, for example, McCullagh 1987, p. 214). Theorem 2 is then immediate.

In problems with independent and identically distributed variables, the forms of $\tilde{l}$ and J are particularly straightforward:

$$\tilde{l}(\beta) = \mathrm{E}_{\beta_0}(l_1(\beta) - l_1(\beta_0))\exp(l_1(\beta) - l_1(\beta_0)) \tag{2.11}$$

and

$$J(\beta) = -\mathrm{E}_{\beta_0}\ddot{l}_1(\beta)\exp(l_1(\beta) - l_1(\beta_0)). \tag{2.12}$$

2.2. Implementation for symbolic computation

To use the above result to find expressions for the k_ps, one needs tractable expressions for $\tilde{l}$ and J. In Appendix 2, we show that

$$\tilde{l}(\beta) = \lim_{n\to\infty} \frac{1}{n} \sum_{p\geqslant 2} \frac{1}{p!} (\beta - \beta_0)^p$$

$$\times \sum_{q_1+2q_2 \ +\ldots+\ kq_k=p} (q_1+\ldots+q_k) b(q_1, \ldots q_k) \operatorname{cum}(\underbrace{l_n^{(1)}, \ldots, l_n^{(1)}}_{q_1 \text{ times}}, \ldots, \underbrace{l_n^{(k)}, \ldots, l_n^{(k)}}_{q_k \text{ times}}), \quad (2.13)$$

where the bs are the coefficients in the Bartlett identities, that is

$$b(q_1, \ldots, q_k) = \frac{p!}{\prod_{i=1}^{k} (i!)^{q_i} q_i!} \quad (2.14)$$

(see, for example, Barndorff-Nielsen and Cox 1989, p. 159). Similarly,

$$J(\beta) = \lim_{n\to\infty} -\frac{1}{n} \sum_{p\geqslant 2} \frac{1}{p!} (\beta - \beta_0)^p$$

$$\times \sum_{q_1+2q_2+\ldots+kq_k=p+2} \tilde{b}(q_1, \ldots, q_k) \operatorname{cum}(\underbrace{l_n^{(1)}, \ldots, l_n^{(1)}}_{q_1 \text{ times}}, \ldots, \underbrace{l_n^{(k)}, \ldots, l_n^{(k)}}_{q_k \text{ times}}), \quad (2.15)$$

where

$$\tilde{b}(q_1, \ldots, q_k) = b(q_1, \ldots, q_k) \frac{1}{\binom{p+2}{2}} \sum_{r=2}^{p+2} \binom{r}{2} q_r. \quad (2.16)$$

Finding the expression for the function (2.10), therefore, is purely a matter of inverting the function $h \to \beta$, and then plugging it into $J(\beta)^{1/2}$ and also differentiating it. This is easily done by symbolic manipulation software; we have used Maple (Char *et al.* 1991) to obtain (2.17) and (2.18) below.

The quantities k_1 and k_2 are both well documented in the literature (see, for example, McCullagh 1987, p. 214). Here, we therefore give

$$k_3 = c_{11}^{-9/2}[- c_{111}c_{11}c_{22} + \tfrac{17}{4} c_{111}c_{11}c_{112} + \tfrac{7}{4} c_{111}c_{11}c_{1111} - \tfrac{125}{72} c_{111}^3 + c_{11}^2 c_{23} - \tfrac{1}{2} c_{11}^2 c_{113}$$
$$- \tfrac{3}{2} c_{11}^2 c_{1112} - \tfrac{3}{10} c_{11}^2 c_{11111}] + O(n^{-1}) \quad (2.17)$$

and

$$k_4 = c_{11}^{-6}[\tfrac{45}{4}c_{11}^2c_{112}c_{1111} - \tfrac{23}{3}c_{111}c_{11}^2c_{23} - \tfrac{9}{8}c_{11}^2c_{22}c_{1111} - \tfrac{45}{4}c_{11}^2c_{22}c_{112} + \tfrac{1465}{144}c_{111}^4 - \tfrac{9}{2}c_{11}^2c_{22}^2$$
$$+ \tfrac{45}{4}c_{11}^2c_{112}^2 + \tfrac{33}{14}c_{11}^2c_{1111}^2 - 6c_{11}^3c_{24} - 8c_{11}^3c_{114} - \tfrac{11}{3}c_{11}^3c_{33} - 12c_{11}^3c_{1113} - \tfrac{51}{2}c_{11}^3c_{1122}$$
$$- \tfrac{21}{2}c_{11}^3c_{11112} - \tfrac{13}{2}c_{11}^3c_{222} - \tfrac{3}{4}c_{11}^3c_{111111} + \tfrac{113}{24}c_{111}^2c_{11}c_{22} - \tfrac{455}{12}c_{111}^2c_{11}c_{112} - \tfrac{341}{24}c_{111}^2c_{11}c_{1111}$$
$$+ \tfrac{19}{3}c_{111}c_{11}^2c_{113} + 3c_{111}c_{11}^2c_{122} + 16c_{111}c_{11}^2c_{1112} + 3c_{111}c_{11}^2c_{11111} - 8c_{11}^3c_{123}] + O(n^{-1}), \tag{2.18}$$

where $c_{q_1\ldots q_r} = \mathrm{cum}(l_n^{(q_1)}, \ldots, l_n^{(q_r)})/n$, and where we have adopted the convention from McCullagh (1987, Section 7.2.3) of using a parametrization where $c_{1q} = 0$ for $q \geqslant 2$. Note that $c_{q_1\ldots q_r}$ depends on n, whereas the k_p do not, which is why there are $O(n^{-1})$ terms in (2.17) and (2.18). We use this to be consistent with (3.3)–(3.5).

One would think of the $c_{q_1\ldots q_r}$ as being of order $O(1)$. In the case of i.i.d. observations, these coefficients are simply constant, and there is no $O(n^{-1})$ term. The more general notation can accommodate triangular arrays, independent but non-identically distributed observations, and also some cases of dependent observations. Conditions under which the $c_{q_1\ldots q_r}$ are $O(1)$ in the dependent case can, for example, be found in Goetze and Hipp (1983).

3. The accuracy of confidence intervals

One of the least studied phenomena of likelihood theory is the impact on the coverage accuracy of confidence intervals of the alternative implied by the likelihood family that is used.

From a traditional likelihood perspective, this may seem like a strange consideration, as the likelihood is determined by the actual family of alternatives. Recent years, however, have seen the increasing use of 'artificial' likelihoods that are designed to work under a multiplicity of null distributions, and such likelihoods need a pragmatic and sometimes deliberately incorrect specification of the family of alternatives. Examples of this include the projective (McLeish and Small 1992), dual (Mykland 1995a; Kong and Cox 1996) and exponential (Nicolae 1999) likelihoods.

The usual set-up is the following. One has a class $\mathscr{P}$ of null distributions P, and one has a score function which we shall write as $\dot{l}_{\mathrm{true},n}(\beta_0)$. The latter notation is a little abusive, as we do not assume that one has a likelihood $l_{\mathrm{true},n}(\beta)$. In fact, in the types of problems we are considering, one would seek to avoid specifying such a likelihood. A main reason would be that such a specification might lead to additional constraints on the class $\mathscr{P}$. To create an R statistic to test $\mathrm{H}_0 : P \in \mathscr{P}$, one specifies instead a criterion function $l_n(\beta)$ satisfying (i) $\mathrm{E}_P \exp\{l_n(\beta) - l_n(\beta_0)\} = 1$ for all $P \in \mathscr{P}$ and (ii) $\dot{l}_n(\beta_0) = \dot{l}_{\mathrm{true},n}(\beta_0)$. This specifies a parametric family $\mathrm{d}P_\beta = \exp\{l_n(\beta) - l_n(\beta_0)\}\mathrm{d}P$ for each $P \in \mathscr{P}$, but the likelihood and the R statistic are independent of which of these families is actually used.

Example. Consider an AR(1) model $X_{n+1} = \beta X_n + \epsilon_{n+1} (|\beta| < 1)$ where one does not know the distribution of the ϵ_ns apart from assuming that they are i.i.d. with mean zero. A frequently used score function for β would be $\dot{l}_{\text{true},n}(\beta) = \sum_{i=0}^{n-1} X_i (X_{i+1} - \beta X_i)$. This is, of course, the likelihood score for Gaussian ϵs with variance 1, as well as the quasi-score (see, for example, Section 9.4 of McCullagh and Nelder 1989, p. 341). There is, however, no likelihood resulting in this score that will generate a bigger family $\mathscr{P}$ than the Gaussian distributions. There are, however, likelihoods, such as the projective or dual likelihood, that will give rise to the score $\dot{l}_n(\beta_0)$ across a bigger $\mathscr{P}$ if one does not require the criterion function to generate all the distributions in $\mathscr{P}$, and if one lets the likelihood depend on β_0. Such likelihoods satisfy (i) and (ii) above, and can be used to create an R statistic.

Requirement (ii) implies that the artificial and the (notional) true likelihoods have the same power to first order in contiguous neighbourhoods (this is fairly obvious, but for a detailed argument in the dual likelihood case, see Mykland 1995a, Section 5); the argument generalizes easily to other situations), and hence, typically, also to second order (Bickel *et al.* 1981). At the third order, though the two likelihoods do not have the same power, there is no clear ordering between them (for a comparison of dual and true likelihood see Lazar and Mykland 1998; the same arguments apply quite generally).

Since the artificial likelihoods, therefore, have good efficiency properties, a comparison of likelihoods will mainly come down to the question of accuracy. This is what we shall pursue in the following. Note that the debate concerning this matter has been particularly acute in connection with empirical/dual likelihood; see Corcoran *et al*. (1995) and Mykland (1999). Incidentally, in view of the efficiency properties of the artificial likelihoods, it may even be relevant to consider this accuracy question when $\mathscr{P}$ has only one element.

The formulae in the previous section permit us to characterize the impact on accuracy of the choice of artificial likelihood. We are looking at log-likelihoods l_n having the same score $\dot{l}_n$, but where we can otherwise vary $\ddot{l}_n$, $\dddot{l}_n$, and so on, as we see fit, so long as we still have a likelihood function. The only restriction we impose is that

$$\operatorname{cov}(\dot{l}_n, \ddot{l}_n) = \operatorname{cov}(\ddot{l}_n, \dddot{l}_n) = \ldots = 0$$

(as in McCullagh 1987, Section 7.2.3), since this can be done by a reparametrization which does not alter the statistic R_n.

In view of property (1.1), the Edgeworth expansion for the cumulative distribution F_n of the signed square root R_n is

$$F_n(r) = \Phi(r) - \phi(r)\left\{ n^{-1/2} k_1 + n^{-1} r \frac{1}{2!}(k_2 + k_1^2) + n^{-3/2} k_1' + n^{-3/2}(r^2 - 1)\left(\frac{1}{3!}(k_3 + k_1^2) + \frac{1}{2!} k_2 k_1\right)\right\} + O(n^{-2}).$$

In the same notation as (2.17)–(2.18), we have that

$$k_1 = -c_{11}^{-3/2} c_{111}/3!, \tag{3.3}$$

$$k_1' - = \bar{k}_1' - c_{11}^{-5/2}[\tfrac{1}{4}c_{23} + \tfrac{1}{4}c_{1112} + \tfrac{1}{12}c_{113}] - c_{11}^{-7/2}c_{111}[\tfrac{17}{24}c_{112} + \tfrac{9}{16}c_{22}] \tag{3.4}$$

and

$$k_2 + c_{11}^{-2}[\tfrac{1}{4}c_{22} - \tfrac{1}{2}c_{112} - \tfrac{1}{4}c_{1111}] + \tfrac{7}{18}c_{11}^{-3}c_{111}^2 + O(n^{-1}), \tag{3.5}$$

where $\bar{k}_1'$ is the corresponding quantity for an exponential family with the same score, that is, $\bar{k}_1' = -(625c_{111}^3 - 630c_{111}c_{1111}c_{11} + 108c_{11111}c_{11}^2)c_{11}^{-9/2}/360$. k_1 and k_2 come from McCullagh (1987, p. 214); (3.4) is derived by a higher-order version of the technology presented in this paper. One could also, obviously, expand the cs in orders of n, but that would only increase the messiness of expressions.

In view of (2.17)–(2.18) and (3.3)–(3.5), the coverage error at the $n^{-1/2}$ level is fixed by the score $\dot{l}_n$, the n^{-1} behaviour depends on the score and $\ddot{l}_n$, the $n^{-3/2}$ behaviour on $\dot{l}_n$, $\ddot{l}_n$ and $\dddot{l}_n$, and so on. This, in itself, is not particularly surprising.

What is surprising, however, is that there is a radical difference between what can go wrong at the n^{-1} level and the $n^{-3/2}$ level. We shall argue below that a bad choice of $\ddot{l}_n$ can result in arbitrary undercoverage, but limited overcoverage. On the other hand, a bad choice of $\dddot{l}_n$ can lead to both unlimited under- and overcoverage.

Consider first the n^{-1} level, that is, the coefficient k_2. We shall argue that the worst-case scenario for overcoverage occurs when the second derivative of the log-likelihood is of the following 'most conservative' form:

$$\ddot{l}_{n,mc} = [\dot{l},\, \dot{l}]_n - 2\,\mathrm{var}(\dot{l}_n) + a_n\dot{l}_n, \tag{3.6}$$

where $[\dot{l},\, \dot{l}]_n$ is the observed (optional) quadratic variation of $\dot{l}_n$. Any non-random a_n will do (one obtains the same R_n statistic). For the validity of our formulae, we need (3.1), and so we assume $a_n = -\mathrm{cov}(\dot{l}_n, [\dot{l},\, \dot{l}]_n)/\mathrm{var}(\dot{l}_n)$. Note that one valid log-likelihood function with this second derivative is simply $\exp\{(\beta - \beta_0)\dot{l}_n + \frac{1}{2}(\beta - \beta_0)^2\ddot{l}_{n,mc}\}$ normalized by its expectation.

To analyse the relationship between the most conservative choice (3.6) and any other second derivative $\ddot{l}_n$, suppose that $\ddot{l}_n = \ddot{l}_{n,mc} + m_n + r_n$, where m_n is a martingale orthogonal to $\dot{l}_n$, and r_n is $O_p(1)$ and asymptotically independent of $\dot{l}_n$, $[\dot{l},\, \dot{l}]_n$ and m_n. This will be the case in most regular situations: the independent case is obvious; for Markov chains, see Jacod and Shiryaev (1987, pp. 445–448); for mixing sums, see Hall and Heyde (1980, Chapter 5), or Jacod and Shiryaev (1987, pp. 448–458). By the Bartlett identities for martingales (Mykland 1994),

$$\mathrm{cov}(\ddot{l}_{n,mc},\, m_n) = \mathrm{cum}(\dot{l}_n,\, \dot{l}_n,\, m_n), \tag{3.7}$$

(use (2.16) in that paper as well as $\mathrm{cov}(\dot{l}_n, [m,\, \dot{l}]_n) = \mathrm{E}([\dot{l},\, \dot{l},\, m]_n)$. Hence

$$k_2 = k_{2,mc} + \frac{1}{4}c_{11}^{-2}\frac{1}{n}\,\mathrm{var}(m_n) + o(1), \tag{3.8}$$

where $k_{2,mc}$ is the value of k_2 when $\ddot{l}_{n,mc}$ is taken as the second derivative of l. Thus,

$$k_2 \geqslant k_{2,mc}, \tag{3.9}$$

As discussed after formula (2.18), c_{11}^{-2} is of order $O(1)$, and so is $\mathrm{var}(m_n)/n = \mathrm{E}[m,\, m]_n/n$ in

asymptotically ergodic situations. Hence, the second term of (3.8) is of order $O(1)$. Similarly, $k_{2,mc} = O(1)$ by (3.5).

Equation (3.9) establishes our claim about limited overcoverage at this level, as follows. For fixed $\dot{l}_n$, let $F_{n,mc}$ be given as the expansion (3.2) up to and including terms of order $O(n^{-1})$ for an R statistic that comes from a log-likelihood satisfying (3.6). This uniquely defines $F_{n,mc}$. In view of the above, we have the following theorem.

Theorem 3. *Subject to regularity conditions, for fixed* $\dot{l}_n$,

$$\begin{aligned} P(R_n \leqslant r) &= F_{n,mc}(r) - (k_2 - k_{2,mc})\frac{1}{2!}n^{-1}r\phi(r) + O(n^{-3/2}) \\ &\leqslant F_{n,mc}(r) + o(n^{-1}), \qquad \textit{for } r > 0, \\ &\geqslant F_{n,mc}(r) + o(n^{-1}), \qquad \textit{for } r < 0. \end{aligned} \tag{3.10}$$

Hence, there is limited overcoverage both for one-sided confidence sets $\{R_n \leqslant r\}$ and $\{R_n \geqslant -r\}$, and for two-sided sets $\{|R_n| \leqslant r\}$, $r > 0$.

The coefficients in the $n^{-3/2}$ term, however, tell a different story. From (3.2), we can write

$$\begin{aligned} P(R_n \leqslant r) = -n^{-3/2}\phi(r)\left\{ k_1' + \frac{1}{3!}(r^2 - 1)k_3 \right\} \\ + \text{terms that only depend on } \dot{l}_n \text{ and } \ddot{l}_n + O(n^{-2}). \end{aligned} \tag{3.11}$$

In both k_3 and k_1', $\ddot{l}_n$ enters linearly. Let us focus on k_3, let $\dot{l}_n$ and $\ddot{l}_n$ be given, and consider a zero-mean martingale m_n, orthogonal to $\dot{l}_n$, so that

$$\operatorname{cov}(\ddot{l}_n,\, m_n) - \tfrac{1}{2}\operatorname{cum}(\dot{l}_n,\, \dot{l}_n,\, m_n) = \nu n + o(n), \tag{3.12}$$

where $\nu \neq 0$. Replace the original $\ddot{l}_n$ by $\ddot{l}_{\alpha,n} = \ddot{l}_n + \alpha m_n$. The new $\ddot{l}_{\alpha,n}$ satisifes the third Bartlett identity (and is hence a valid third derivative of l_n), and also $\operatorname{cov}(\ddot{l}_{\alpha,n},\, \dot{l}_n) = 0$. In this set-up,

$$k_{3,\alpha} = k_3 + \alpha c_{11}^{-5/2}\nu + o(1), \tag{3.13}$$

which can take any value. In other words, in view of (3.11), both under- and overcoverage are potentially unbounded at this level.

Acknowledgements

This research was supported in part by National Science Foundation grants DMS 96-26266 and DMS 99-71738 and Army Research Office grant DAAH04-95-1-0105. The paper was prepared using computer facilities supported in part by the National Science Foundation grants DMS 89-05292 and DMS 87-03942 awarded to the Department of Statistics at the University of Chicago, and by The University of Chicago Block Fund.

I would like to thank Ole Barndorff-Nielsen, Jens Jensen, Peter McCullagh, Nicole Lazar,

Ib Skovgaard, Jamie Stafford, Trevor Sweeting and Andy Wood for extremely useful discussions. I would also like to thank the referee for a careful reading and an extremely useful report. Thanks also to Mitzi Nakatsuka for typing part of the paper.

Appendix 1: Curved exponential families and proof of Theorem 1

For a more rigorous development, consider a curved exponential family

$$l_n(\beta) = l_n(\beta_0) + (\beta - \beta_0)\dot{l}_n(\beta_0) + \tfrac{1}{2}(\beta - \beta_0)^2 \ddot{l}_n(\beta_0) + \dots \tag{A1.1}$$

of order p (i.e., terms of order $p+1$ and higher are non-random). We shall consider R for testing $H_0 : \beta = \beta_0$. Suppose that there is a valid saddlepoint approximation to the density of the vector $(\dot{l}_n(\beta), \dots, l_n^{(p)}(\beta))$. One can then proceed as follows.

Begin by fixing $\beta_1 \neq \beta_0$. Then reparametrize the family as in Section 7.2.3 of McCullagh (1987, pp. 204–207) to make $\mathrm{cov}_{\beta_1}(\dot{l}_n(\beta_1), l_n^{(q)}(\beta_1)) = 0$ for $2 \leqslant q \leqslant p$. From McCullagh (1987), this is accomplished by using the parameter ϕ, given by $\phi_1 = \beta$, and

$$\begin{aligned}\phi(\beta) - \phi_1 &= \beta - \beta_1 + \frac{1}{2}(\beta - \beta_1)^2 \frac{\mathrm{cov}_{\beta_1}(\dot{l}_n(\beta_1), \ddot{l}_n(\beta_1))}{\mathrm{var}_{\beta_1}(\dot{l}_n(\beta_1))} + \dots \\ &= \frac{\mathrm{E}_{\beta_1} \dot{l}_n(\beta_1)(l_n(\beta) - l_n(\beta_1))}{\mathrm{var}_{\beta_1}(\dot{l}_n(\beta_1))}.\end{aligned} \tag{A1.2}$$

Hence

$$\begin{aligned}\phi_0 - \phi_1 &= \phi(\beta_0) - \phi_1 \\ &= \frac{\dfrac{\partial}{\partial \beta} \mathrm{E}_{\beta_0} g(l_n(\beta) - l_n(\beta_0))|_{\beta=\beta_1}}{\mathrm{E}_{\beta_1} \ddot{l}_n(\beta_1)} \\ &\approx -\frac{\dot{\dot{l}}(\beta_1)}{J(\beta_1)}\end{aligned} \tag{A1.3}$$

as $n \to \infty$ under P_β. Here $g(x) = (x-1)\mathrm{e}^x$, which can be replaced by $g(x) = x\mathrm{e}^x$ since $\mathrm{E}_{\beta_0} \exp(l_n(\beta_1) - l_n(\beta_0)) = 1$. In the new parametrization, the null hypothesis is $\phi = \phi_0$.

Now embed $l_n(\beta) - l_n(\beta_0)$ in a full exponential family. In the notation of Jensen (1997) (which we shall be using in the following), $\overline{T}_q = l_n^{(q)}/q!n$. Note that we do not require the $\overline{T}_q$s to be means, only that the saddlepoint approximation hold.

Our larger family is then (in the new parametrization)

$$l_n(\phi_0) + \theta_1 \dot{l}_n(\phi_0) + \dots + \theta_p \frac{1}{p!} l_n^{(p)}(\phi_0).$$

A reparametrization of the θs is given by

$$\theta_1 = \phi$$

and

$$\theta_l = \phi b_l, \qquad \text{for } l \geqslant 2 \tag{A1.4}$$

(in Jensen's notation, ϕ is β_0 and b_l is β_l). A corresponding sequence of null hypotheses is

$$H_0^{(1)}: \phi = \phi_0$$

and

$$H_0^{(l)}: b_l = 1, \qquad \text{for } l \geqslant 2. \tag{A1.5}$$

Hence, $H_0^{(1)}$ is our original null hypothesis.

Let the $\bar{u}_l$s be chosen as in Jensen (1997, Section 3). In view of Section 2 of the same paper, the joint density of $(R_1, R_{L,2}, \ldots, R_{L,p})$ is, in a large-deviation region,

$$\frac{1}{(2\pi)^{p/2}} \frac{h_1}{\bar{u}_1} \exp\left(-\frac{1}{2} r_1^2 - \frac{1}{2} \sum_{i=2}^{p} r_{l,i}^2 \right) \{1 + O(n^{-1})\}. \tag{A1.6}$$

We are here suppressing the dependence of the Rs on n, and similarly for the Us and $\hat{\theta}$s below. Note that $R_1 = R$. By using Skorokhod embedding, it therefore follows that, under P_{β_1},

$$\frac{f_{\beta_0,n}(R|R_{L,2}, \ldots, R_{L,p})}{\phi(R)} = \frac{R_1}{\overline{U}_1} \{1 + O_p(n^{-1})\}. \tag{A1.7}$$

Note that $R_1/\sqrt{n} = h(\beta_1)(1 + O_p(n^{-1/2}))$, where h is as given in (2.8). Hence, if we can show that

$$\overline{U}_1/\sqrt{n} = (\phi_1 - \phi_0) J(\beta_1)^{1/2} \{1 + O_p(n^{-1/2})\}, \tag{A1.8}$$

it follows from (A1.3), and by averaging over $(R_{L,2}, \ldots, R_{L,p})$, that

$$\begin{aligned} \frac{f_{\beta_0,n}(R)}{\phi(R)} &= \frac{J(\beta_1)^{1/2} h(\beta_1)}{\dot{\bar{l}}(\beta_1)} \{1 + O_p(n^{-1/2})\} \\ &= J^{1/2}(\beta_1) \frac{\partial \beta}{\partial h}(h) \{1 + O_p(n^{-1/2})\} \end{aligned} \tag{A1.9}$$

under P_{β_1}. Again by Skorokhod embedding, we obtain Theorem 1.

It remains to show (A1.8). In Jensen's (1997) notation,

$$\begin{aligned} \hat{\theta}^l - \hat{\theta}^{l-1} = (&\hat{\beta}_1^l - \hat{\beta}_1^{l-1}, \hat{\theta}_2^l - \hat{\theta}_2^{l-1}, \ldots, \hat{\theta}_{l-1}^l - \hat{\theta}_{l-1}^{l-1}, \hat{\theta}_l^l - (\hat{\beta}_1^{l-1})^l, \\ &(\hat{\beta}_1^l) - (\hat{\beta}_1^{l-1})^{l+1}, \ldots, (\hat{\beta}_1^l)^p - (\hat{\beta}_1^{l-1})^p) \end{aligned} \tag{A1.10}$$

where $\hat{\theta}_l - (\hat{\beta}_1^{l-1})^l$ is the term in the lth column. Note that $(\hat{\beta}_1^l)^k$ is β_1^l raised to power k, which is the only instance of power notation in (A1.10).

Since $\text{corr}(\overline{T}_1, \overline{T}_l) \simeq 0$ (note that this is where the reparametrization above is used),

$\hat{\theta}_l^l - (\hat{\beta}_1^{l-1})^l$ is $O_p(n^{-1/2})$ but not $o_p(1)$. On the other hand, for $l \geqslant 2$, $\hat{\beta}_1^l - \hat{\beta}_1^{l-1} = O_p(n^{-1})$. Hence, for $l \geqslant 2$,

$$\hat{\theta}^l - \hat{\theta}^{l-1} = (0, \hat{\theta}_2^l - \hat{\theta}_2^{l-1}, \ldots, \hat{\theta}_l^l - (\hat{\beta}_1^{l-1})^l, 0, \ldots, 0) + O_p(n^{-1}). \tag{A1.11}$$

Hence the determinants in equation (7) in Jensen (1997) can be evaluated by multiplying the diagonal, and so (5.8) follows. Note that in the above argument, if $\overline{T}_l$ is zero, one just deletes line and column l and makes the appropriate modification to the next column. This does not affect the result.

Appendix 2: Derivations for Section 2.2

If $g(x) = x\mathrm{e}^x$,

$$\begin{aligned}\tilde{l}(\beta) &= \lim_{n\to\infty} \frac{1}{n} \mathrm{E}_\beta (l_n(\beta) - l_n(\beta_0)) \\ &= \lim_{n\to\infty} \frac{1}{n} \mathrm{E}_{\beta_0} g(l_n(\beta) - l_n(\beta_0)),\end{aligned} \tag{A2.1}$$

so that the pth derivative is

$$\tilde{l}^{(p)}(\beta_0) = \lim_{n\to\infty} \frac{1}{n} \mathrm{E}_{\beta_0} \sum_{q_1+2q_2+\cdots+kq_k=p} g^{(q_1+\cdots+q_k)}(0) b(q_1, \ldots, q_k) \dot{l}_n(\beta_0)^{q_1} \cdots l_n^{(k)}(\beta_0)^{q_k}, \tag{A2.2}$$

which yields (2.13) since $g^{(v)}(0) = v$ and since moments can be replaced by cumulants in the above. The latter can either be seen by direct computation, or by observing that the right-hand side of (A2.1) is $O(1)$.

To find J, note that

$$\begin{aligned}&\frac{\partial^p}{\partial \beta^p} \ddot{l}_n(\beta) \exp(l_n(\beta) - l_n(\beta_0))|_{\beta=\beta_0} \\ &\quad = \sum_{r=0}^{p} \binom{p}{r} l_n^{(r+2)} \sum_{q_1+2q_2+\ldots+kq_k=p-r} b(q_1, \ldots, q_v)(\dot{l}_n)^{q_1} \ldots (l_n^{(k)})^{q_k},\end{aligned} \tag{A2.3}$$

which gives (2.15) for the same reasons as used above, and because

$$\tilde{b}(q_1, \ldots, q_v) = \sum_{r=0}^{p} \binom{p}{r} b(q_1, \ldots, q_{r+1}, q_{r+2} - 1, q_{r+3}, \ldots, q_v). \tag{A2.4}$$

Since

$$b(q_1, \ldots, q_{r+1}, q_{r+2} - 1, q_{r+3}, \ldots, q_v) = b(q_1, \ldots, q_v) \frac{q_{r+2}}{\binom{p+2}{r+2}}, \tag{A2.5}$$

this gives (2.15) by direct computation.

References

Andrews, D.F. and Stafford, J.E. (1993) Tools for the symbolic computation of asymptotic expansions. *J. Roy. Statist. Soc. Ser. B*, **55**, 613–627.

Andrews, D.F. and Stafford, J.E. (2000) *Symbolic Computation for Statistical Inference*. Oxford: Oxford University Press.

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

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

Barndorff-Nielsen, O.E. and Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics*. London: Chapman & Hall.

Barndorff-Nielsen, O.E. and Wood, A.T.A. (1998) On large deviations and choice of ancillary for p^* and r^*. *Bernoulli*, **4**, 35–63.

Bartlett, M.S. (1953a) Approximate confidence intervals. *Biometrika*, **40**, 12–19.

Bartlett, M.S. (1953b) Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, **40**, 306–317.

Bhattacharya, R.N. and Ghosh, J.K. (1978) On the validity of the formal Edgeworth expansion. *Ann. Statist.*, **6**, 434–451.

Bickel, P.J., Chibisov, D.M. and van Zwet, W.R. (1981) On efficiency of first and second order. *Internat. Statist. Rev.*, **49**, 169–175.

Chaganty, N.R. and Sethuraman, J. (1985) Large deviation local limit theorems for arbitrary sequences of random variables. *Ann. Probab.*, **13**, 97–114.

Char, B.W., Geddes, K.O., Gonnet, G.H., Leong, B.L., Monagan, M.B. and Watt, S.M. (1991) *Maple V. Language Reference Manual*. Berlin: Springer-Verlag.

Corcoran, S.A., Davison, A.C. and Spady, R.H. (1995) Reliable inference from empirical likelihoods. Preprint.

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

DiCiccio, T.J. and Romano, J.P. (1989) On adjustments based on the signed root of the empirical likelihood ratio statistic. *Biometrika*, **76**, 447–456.

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

DiCiccio, T.J. and Stern, S.E. (1994) Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihoods. *J. Roy. Statist. Soc. Ser. B*, **56**, 397–408.

DiCiccio, T.J., Hall P. and Romano, J.P. (1991) Empirical likelihood is Bartlett-correctable. *Ann. Statist.*, **19**, 1053–1061.

Goetze, F. and Hipp, C. (1983) Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **64**, 211–239.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion.* Berlin: Springer-Verlag.

Hall, P. and Heyde, C.C. (1980) *Martingale Limit Theory and its Application.* New York: Academic Press.

Jacod, J. and Shiryaev, A.N. (1987) *Limit Theorems for Stochastic Processes*. Berlin: Springer-Verlag.

Jensen, J.L. (1992) The modified signed likelihood statistic and saddlepoint approximations. *Biometrika*, **79**, 693–703.

Jensen, J.L. (1995) *Saddlepoint Approximations in Statistics*. Oxford: Oxford University Press.

Jensen, J.L. (1997) A simple derivation of r^* for curved exponential families. *Scand. J. Statist.*, **24**, 33–46.

Kong, A. and Cox, N.J. (1996) From efficient nonparametric tests for linkage analysis to

semiparametric models and lodscores. Technical report no. 435, Department of Statistics, University of Chicago.
Lawley, D.N. (1956) A general method for approximating the distribution of likelihood ratio criteria. *Biometrika*, **43**, 295–303.
Lazar, N. and Mykland, P.A. (1998) An evaluation of the power and conditionality properties of empirical likelihood. *Biometrika*, **85**, 523–534.
McCullagh, P. (1984) Local sufficiency. *Biometrika*, **71**, 233–244.
McCullagh, P. (1987) *Tensor Methods in Statistics*. London: Chapman & Hall.
McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edition. London: Chapman & Hall.
McCullagh, P. and Tibshirani, R. (1990) A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B*, **52**, 325–344.
McLeish, D.L. and Small, C.G. (1992) A projected likelihood function for semiparametric models. *Biometrika*, **79**, 93–102.
Mykland, P.A. (1994) Bartlett type identities for martingales. *Ann. Statist.*, **22**, 21–38.
Mykland, P.A. (1995a) Dual likelihood. *Ann. Statist.*, **23**, 396–421.
Mykland, P.A. (1995b) Embedding and asymptotic expansions for martingales. *Probab. Theory Related Fields*, **103**, 475–492.
Mykland, P.A. (1999) Bartlett identities and large deviations in likelihood theory. *Ann. Statist.*, **27**, 1105–1117.
Nicolae, D.L. (1999) Allele sharing models in gene mapping: a likelihood approach. Ph.D. thesis, University of Chicago.
Peers, H.W. and Iqbal, M. (1985) Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications, *J. Roy. Statist. Soc. Ser. B*, **47**, 547–554.
Shenton, L.R. and Bowman, K.O. (1977) *Maximum Likelihood Estimation in Small Samples*. London: Griffin.
Skovgaard, Ib (1986) A note on the differentiation of cumulants of log likelihood derivatives. *Internat. Statist. Rev.*, **54**, 29–32.
Skovgaard, Ib (1990) On the density of minimum contrast estimators. *Ann. Statist.*, **18**, 779–789.
Skovgaard, Ib (1996) An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, **2**, 145–165.
Stafford, J.E. (1994) Automating the partition of indexes. *J. Comput. Graph. Statist.*, **3**, 249–259.
Stafford, J.E. (1995) Exact cumulant calculations for Pearson χ^2 and Zelterman statistics for r-way contingency tables, *J. Comput. Graph. Statist.*, **4**, 199–212.
Stafford, J.E. and Andrews, D.F. (1993) A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, **80**, 715–730.
Stafford, J.E., Andrews, D.F. and Wang. Y. (1994) Symbolic computation: a unified approach to studying likelihood. *Statist. Comput.*, **4**, 235–245.
Wallace, D.L. (1958) Asymptotic approximations to distributions. *Ann. Math. Statist.*, **29**, 635–654.

Received August 1998 and revised January 2001