The Annals of Statistics
1999, Vol. 27, No. 3, 1105–1117

BARTLETT IDENTITIES AND LARGE DEVIATIONS IN LIKELIHOOD THEORY[1]

BY PER ASLAK MYKLAND

University of Chicago

The connection between large and small deviation results for the signed square root statistic R is studied, both for likelihoods and for likelihood-like criterion functions. We show that if $p-1$ Barlett identities are satisfied to first order, but the pth identity is violated to this order, then $\mathrm{cum}_q(R) = O(n^{-q/2})$ for $3 \leq q < p$, whereas $\mathrm{cum}_p(R) = \kappa_p n^{-(p-2)/2} + O(n^{-p/2})$. We also show that the large deviation behavior of R is determined by the values of p and κ_p. The latter result is also valid for more general statistics. Affine (additive and/or multiplicative) correction to R and R^2 are special cases corresponding to $p = 3$ and 4. The cumulant behavior of R gives a way of characterizing the extent to which R-statistics derived from criterion functions other than log likelihoods can be expected to behave like ones derived from true log likelihoods, by looking at the number of Bartlett identities that are satisfied. Empirical and nonparametric survival analysis type likelihoods are analyzed from this perspective via the device of "dual criterion functions."

1. Introduction. What does it take for a criterion function to resemble a likelihood? This is gradually becoming an important question with the proliferation of approximate likelihoods. By this we mean criterion functions l for which $E_{\theta_0} \exp(l(\theta) - l(\theta_0))$ is approximately 1, but not exactly so. For this purpose, ordinary, partial [Cox (1975), Wong (1986)], projective [McLeish and Small (1992)] and dual [Mykland (1995)] likelihoods are exact, whereas the following are approximate:

1. Quasi-likelihood [Wedderburn (1974), Godambe and Heyde (1987); see also McCullagh and Nelder (1989); cf. the end of Section 3];
2. Empirical likelihood [Owen (1988), exact for means and estimating equations; cf. Section 4];
3. Nonparametric point process likelihood [Jacod (1975); see also Andersen, Borgan, Gill and Keiding (1993); cf. Section 4];
4. Sieve likelihood [Wong and Severini (1991)] and
5. Bootstrap likelihood [Davison, Hinkley and Worton (1992)].

Received March 1997; revised February 15, 1999.

[1] Supported in part by NSF Grant DMS-96-26266 and Army Research Office grant DAAH 04-95-1-0105.

AMS 1991 *subject classifications*. Primary 62E20, 62M09, 62M10; secondary 60G42, 60G44, 62J99, 62M99.

Key words and phrases. Approximate likelihoods, Bartlett correction, convergence of cumulants, dual criterion functions, empirical likelihood, survival analysis.

This is in no way an exhaustive list; one could include, for example, profile likelihoods with various adjustments, such as those of Barndorff-Nielsen (1983), Cox and Reid (1987) and McCullagh and Tibshirani (1990).

There are, obviously, several aspects to this question. We shall in the following be concerned with accuracy of asymptotic approximations and specifically with the likelihood ratio statistic R^2 and its signed square root R. Other questions that need to be asked include efficiency and what is often called inferential correctness, but we shall not discuss these issues here.

The accuracy question is a very important one. The approximation properties of R and R^2, either taken "raw" or corrected in ways to be discussed below, are usually superior to those of most other statistics on offer. The normal and χ^2 approximations are generally excellent in small samples as most simulation studies will confirm, which is not to say that counter-examples cannot be found.

A major question is then what large sample properties, if any, are there that attach to R and R^2 and that can possibly predict their nice behavior in small samples. One approach to this issue is to study Edgeworth expansion properties of the mean and variance corrected R and R^2 [Lawley (1956), Barndorff-Nielsen and Cox (1979, 1984), McCullagh (1987)]. Then R can be adjusted to $(R - E(R))/\mathrm{sd}(R)$, and R^2 to $R^2 \times d/E(R^2)$, where d is the number of degrees of freedom. The latter is the famous Bartlett correction. Such correctability has also been studied for empirical likelihood by DiCiccio and Romano (1989) and DiCiccio, Hall and Romano (1991). These corrections to R and R^2 are often referred to as "affine" corrections, as the new statistic is an affine transformation of the original one.

Affine correctability, however, is by no means the whole story, as witnessed by the substantial body of work on large deviations; see, in particular, Barndorff-Nielsen and Wood (1998), Jensen (1992, 1995, 1997) and Skovgaard (1990, 1996). Starting with Barndorff-Nielsen (1986), this approach also yields types of correction other than the affine ones, such as in the R^* statistic.

We shall show in this paper that these two angles of research can be unified, at least for smooth families. The connection between small and large deviation expansions is, at least heuristically, clear-cut. They both rely on cumulants, though the cumulants show up in terms of different order in the two types of expansion. For an example of a rigorous study concerning this connection, see Robinson, Höglund, Holst and Quine (1990).

Our angle on this is to show a result (in Section 2) describing how the intermediate deviation behavior of the density of a statistic is controlled by the asymptotic behavior of its cumulants. From this, we shall use large deviation properties of R to calculate the first-order structure of $\mathrm{cum}_p(R)$, and we shall see that this yields a remarkable property which generalizes affine correctability.

Our results are then used (in Section 3) to assess how close a criterion function is to a likelihood. We shall see that this relates to the number of Bartlett identities satisfied to first order. As a main application, we shall analyze empirical and survival analysis-type likelihoods through what we call dual criterion functions (Section 4).

2. Small and large deviations. An important folk theorem in statistics is that standard behavior of an asymptotically normal statistic T_n is, for $q \geq 3$,

$$\mathrm{cum}_q(T_n) = n^{-(q-2)/2}\kappa_q + O(n^{-q/2}). \tag{2.1}$$

In the case of statistics that are functions of means of i.i.d. random variables, conditions for this to hold are set out in Theorem 2.1, page 53, of Hall (1992). This asymptotic behavior, however, is valid in considerably greater generality, as discussed in Chapters 2.3 and 2.4 of Hall (1992). The result (2.1) is crucial to the behavior of Edgeworth expansions of asymptotically normal statistics, see, in particular, Wallace (1958), Bhattacharya and Ghosh (1978) and Hall (1992).

What is special about the small deviation behavior of R and R^2 is that, for R, κ_3 and κ_4 are zero. This is what leads the null distribution of the affinely corrected statistics to be $N(0,1) + O(n^{-3/2})$ and $\chi^2 + O(n^{-2})$, respectively, in the sense of Edgeworth and related expansions; see, for example, Chapter 7 of McCullagh (1987). Affine correction, however, cannot take you any further than this; in general, the $O(n^{-3/2})$ term in the Edgeworth expansion for R, even when affinely corrected, does not vanish. This is because the $O(n^{-3/2})$ term in the expansion for $\mathrm{cum}_3(R)$ does not vanish, as we shall see at the end of this section.

What we shall see in the following is that R satisfies

$$\kappa_q = 0 \quad \text{for all } q \geq 3. \tag{2.2}$$

In other words, affine correctability is only a special case, relating to cumulants number 3 and 4, of a property affecting all cumulants. Also, since the $O(n^{-q/2})$ term in the expansion (2.1) does not generally vanish for the R statistic (see the end of this section) (2.2) would seem to be the main asymptotic property governing the accuracy behavior of R.

Why is this important? After all, in an Edgeworth expansion, the $O(n^{-3/2})$ term for the third cumulant is as important as κ_5, and more important than κ_6.

What we shall show (Theorem 1) is that for largeish deviations, this picture is different. In such a setup, the asymptotic behavior of the density of a statistic T_n is characterized by the first κ_q that is nonzero. The bigger this q is, the better off one is approximation-wise. The fact that, for R, $q = +\infty$, may go some way towards explaining its nice small sample properties.

The theory hinges on the saddlepoint approximation. Under a variety of regularity conditions, one can show that for an asymptotically normal statistic T_n with density f_n, one can write

$$f_n(r) = \frac{1}{(2\pi K_n''(\hat{\tau}_n))^{1/2}} \exp\bigl(K_n(\hat{\tau}_n) - \hat{\tau}_n K_n'(\hat{\tau}_n)\bigr)(1 + o(1)), \tag{2.3}$$

where $K_n'(\hat{\tau}_n) = r$. Here, K_n is the cumulant generating function for T_n, or an approximation to this function. The approximation (2.3) is often valid uniformly either for all r, or in a large deviation neighborhood $r\varepsilon[-cn^{\alpha}, cn^{\alpha}]$.

In the statistics literature, such approximations go back to Daniels (1954). An important paper is Chaganty and Sethuraman (1985), where conditions are found for the approximation to hold for a general statistic (and not just a mean). Another way to proceed for general statistics is to adapt the "smooth function of means" model; see in particular Chapter 4 of Jensen (1995). This approach has also been very successful in permitting the establishment of rigorous results for Edgeworth expansions; see, for example, Bhattacharya and Ghosh (1978) and Hall (1992). Note that R-statistics in finite-dimensional curved exponential families are covered by this model.

To see the asymptotic role of of the first nonzero κ_q from (2.1), consider the following result.

THEOREM 1. *Suppose that T_n is a statistic with density f_n and cumulant generating function K_n. Suppose that* (2.3) *holds, uniformly in $r\varepsilon[-cn^{\alpha}, cn^{\alpha}]$, for some $c > 0$ and some α, $1/6 \leq \alpha \leq 1/2$. Suppose, for some integer p, $3 \leq p \leq 1/(0.5-\alpha)$, that the p first cumulants satisfy* (2.1), *with $\kappa_1 = 0$ and $\kappa_2 = 1$. Also suppose that K_n is at least $p+1$ times continuously differentiable, and that $\psi_n^{(p+1)}(t)$ is bounded uniformly in n and $|t| \leq c'n^{-1/p}$, where $c' > 0$ and $\psi_n(t) = K_n(\sqrt{n}t)/n$. The following statements are equivalent*:

(i)

$$\kappa_q = 0 \qquad \textit{for } 2 < q < p; \tag{2.4}$$

(ii) *If r_n is of order $O(n^{1/2-1/p})$ or smaller as n tends to infinity,*

$$f_{T_n}(r_n) = \phi(r_n)\exp\left(\frac{1}{p!}\kappa_p r_n^p n^{-(p-2)/2}\right)(1+o(1)); \tag{2.5}$$

(iii) *If $r_n = sn^{1/2-1/p}$, then*

$$P(T_n \geq r_n) = (1-\Phi(r_n))\exp\left(\frac{1}{p!}\kappa_p r_n^p n^{-(p-2)/2}\right)(1+o(1)), \tag{2.6}$$

uniformly for $s\varepsilon[c_1, c_2]$, where $0 < c_1 < c_2 < \infty$.

In other words, the large deviation properties of T are controlled by κ_p. Note that under the conditions of Theorem 2.1 of Chagany and Sethuraman (1985), the $1+o(1)$ term in (2.5) can be replaced by $1+O(rn^{-1/2})+O(n^{-1})$. This is easily seen from the proof in the Appendix. The result covers the case where K_n is exactly the cumulant generating function. Similar conclusions can, obviously, be obtained in the function-of-means case.

A corollary to this is that if $\alpha = 1/2$, (2.2) is equivalent to

$$f_{T_n}(r) = \phi(r)(1+o(1)) \tag{2.7}$$

and

$$P(T_n \geq r) \ = \ (1-\Phi(r)(1+o(1)), \tag{2.8}$$

uniformly for r of order $O(n^{\alpha})$, for any α, $0 < \alpha < \frac{1}{2}$.

We now return to the signed square root statistic R. One of the main results of contemporary likelihood theory [cf. Barndorff-Nielsen and Wood (1998), Jensen (1992, 1995, 1997) and Skovgaard (1990, 1996)] is that, in curved exponential families, and subject to the existence of all moments and to regularity conditions, for r of order $o(n^{1/2})$, (2.7) and (2.8) do indeed hold when $T_n = R_n$ where R_n is the statistic R based on n observations. One can now use this to infer that (2.2) holds for R statistics from curved exponential families.

In fact, the family does not need to be of the curved exponential type; (2.2) remains valid in the smooth case – analytic families as in Skovgaard (1991). By the type of stochastic expansion given on page 214 in McCullagh (1987), one can approximate R_n by $\overline{R}_n + O_p(n^{-p/2})$, where $\overline{R}_n$ is the corresponding statistic in a $(1, p+2)$ curved exponential family, and hence (2.2) also holds for R_n.

One can then use Theorem 1 to assert that under the conditions of this theorem, (2.7) and (2.8) are also valid for R-statistics from analytic families. We have not investigated the precise regularity conditions needed to assert (2.7) or (2.8) on the assumption that such conditions are not likely to prove particularly informative.

Finally, we provide an example to show that that one cannot, in general, expect the pth cumulant of R to vanish at the $O(n^{-p/2})$ level. Consider $p = 3$, and write

$$\mathrm{cum}_3(R) = n^{-3/2}\nu_3 + O(n^{-5/2}). \tag{2.9}$$

The rationale for focusing on ν_3 might be that since (2.2) is satisfied, ν_3 is the leading error term in the expansion of $f_{R_n}(r)$ for moderate deviations as n tends to infinity, that is, $r \to \infty$, but $r = o(n^{1/2})$.

It is easy to see that for the exponential family $\exp(\theta T - k(\theta))$,

$$\nu_3^{\mathrm{EXPO}} = -\big(625\, k_3^3 - 630 k_3 k_4 k_2 + 108 k_5 k_2^2\big)\frac{1}{360} k_2^{-9/2}, \tag{2.10}$$

where k_q is the qth derivative of $k(\theta)$. Similarly, the expression for ν_3 for empirical likelihood is nonzero and is given by Corcoran, Davison and Spady (1995).

3. Approximate likelihoods. So what happens when a criterion function $l(\theta)$ is not quite a log likelihood?

The connection between Theorem 1 and approximate log likelihood functions can be characterized as follows. We shall here still only be concerned with the one parameter case, and, as before, the statistic R is defined by

$$R = \mathrm{sign}(\hat{\theta} - \theta)\big(2(l(\hat{\theta}) - l(\theta))\big)^{1/2}, \tag{3.1}$$

where $\hat{\theta}$ maximizes $l(\theta)$.

The characterization is now that if all $l^{(q)} = \partial^q l/\partial\theta^q$ exist and all cumulants of the $l^{(q)}$ are of order $O(n)$, then (2.2) remains valid if all the Bartlett identities [Bartlett (1953a, b); see McCullagh (1987)] hold to first order, that is,

$$E(l^{(q)}) + \cdots + \mathrm{cum}_q(\dot{l}) = O(1) \tag{3.2}$$

for all q. If this is not the case, so that (3.2) is only valid for $q < p$ $(p > 2)$, with

$$E(l^{(p)}) + \cdots + \mathrm{cum}_p(\dot{l}) = nk_p + O(1), \tag{3.3}$$

then $\mathrm{cum}_q(R) = O(n^{-q/2})$ for $q < p$, $(q \neq 2)$, and $\mathrm{cum}_p(R) = n^{-(p-2)/2}\kappa_p + O(n^{-p/2})$, with

$$\kappa_p = \frac{1}{p}\sigma^{-p}k_p, \tag{3.4}$$

where σ^2 is asymptotically equal to $\mathrm{var}(\dot{l})/n$. See the Appendix for the derivation.

In particular, affine correctability [to $O(n^{-3/2})$] of R and R^2 depends on (3.2) being valid for $q \leq 4$. As an example of a criterion function that is particularly far from likelihood in the sense that we consider, note that quasi-likelihood only satisfies (3.2) for $q = 1$ and 2.

Since the coefficients in the Taylor expansion for the expectation of the exponential of the criterion function can be expressed as sums of products of the terms on the left-hand side of (3.2) and (3.3) [cf. Example 7.1, page 222, of McCullagh (1987)], (3.2) $(q < p)$ and (3.3) are equivalent to

$$E\exp\bigl(l(\theta) - l(\theta_0)\bigr) = n(\theta - \theta_0)^p k_p \frac{1}{p!} + O\bigl((\theta - \theta_0)^{p+1}\bigr) + u_n(\theta), \tag{3.5}$$

for fixed n, where $u_n(\theta) = O(1)$ as $n \to \infty$. Similarly, (2.2) is, subject to regularity conditions, the same as

$$E\exp\bigl(l(\theta) - l(\theta_0)\bigr) = O(1), \tag{3.6}$$

as $n \to \infty$. Hence, $n(\theta - \theta_0)^p k_p(1/p!) + O((\theta - \theta_0)^{p+1})$ is the measure of deviation of a criterion function from approximate likelihoodness.

It should be emphasized that the cumulant approach outlined above only works in the regular cases covered by our conditions. For an example of an innocuous looking case which falls outside this framework, see Lazar and Mykland (1999).

4. Dual criterion functions.

By this we mean criterion functions which are parametric (i.e., they only depend on finitely many parameters, even as $n \to \infty$) and which give rise to the same R statistic as a corresponding nonparametric likelihood. We shall look at two cases: empirical and nonparametric point process likelihood (the latter as in survival analysis).

The dual criterion function arises as a profile Lagrangian, which becomes a function of the Lagrange multiplier. This multiplier is then the "parameter" in the dual criterion function. In the two cases under study, the dual criterion functions are given by (4.1) and (4.8), respectively. Note that the concept of dual criterion function is left deliberately vague; it is not defined in any rigorous generality. It is intended as a tool for analyzing various nonparametric problems, and it has to be adapted to each individual case. One could, for example, use it to analyze a variety of survival analysis likelihood ratio statistics, such as the one proposed by Thomas and Grunkemeier (1975) based on the Kaplan and Meier (1958) nonparametric likelihood.

The two nonparametric likelihoods we shall study have in common that they are not likelihoods in the sense of being dominated. A likelihood ratio only exists with respect to other probabilities supported on the observations. They can, obviously, be obtained by profiling dominated likelihoods, but that in itself does not assure likelihood behavior in the sense of (2.2), (2.7) and (2.8).

Indeed, as we shall see, the accuracy behavior is not the same for these two types of nonparametric likelihood.

4.1. *Empirical likelihood.* Suppose one observes i.i.d. vectors $X_1, \dots, X_n$ drawn from a distribution F. One is particularly interested in the parameter $\theta = \theta(F)$. The empirical likelihood ratio statistic [Owen (1988)] for testing H_0: $\theta = \theta_0$ is is given by $2(n \log n^{-1} - \hat{l}_E)$, where $\hat{l}_E$ is the maximum of $\log \prod p_i$, subject to $\Sigma p_i = 1$ and $\theta(\hat{F}_p) = \theta_0$. Here, $\hat{F}_p$ is the distribution which puts mass p_i on X_i.

To introduce the dual criterion function, set

$$\omega(\mu, \lambda) = \sup_{p_i} \left\{ \sum_{i=1}^{n} \log p_i - \lambda(\Sigma p_i - 1) - \mu\lambda(\theta(\hat{F}_p) - \theta_0) \right\},$$

where $\omega(\mu, 0)$ is the limit of $\omega(\mu, \lambda)$ as $\lambda \to 0$. Supposing that the supremum is attained, let $\hat{p}_i = \hat{p}_i(\mu, \lambda)$ be the maximizing arguments in the above, and set $\hat{\lambda} = \hat{\lambda}(\mu)$ to be such that $\Sigma \hat{p}_i = 1$. We now define the dual criterion function to be

$$l_D(\mu) = -\omega(\mu, \hat{\lambda}(\mu)). \tag{4.1}$$

A standard optimization argument yields that the resulting R is the same as the R for testing $\mu = 0$ with the dual criterion function $l_D(\mu)$. The problem can therefore be analyzed with the help of the finite parameter methods developed in the preceding sections. We emphasize that the statistic has the usual form, $R = \operatorname{sgn}(\hat{\mu})\sqrt{2(l_D(\hat{\mu}) - l_D(0))}$, where $\hat{\mu}$ is the maximizer, the MLE, of $l_D(\mu)$. The reason why we maximize with respect to μ is that the sign has been changed in (4.1).

If $\theta(\hat{F}_p)$ is a mean, then $l_D(\mu)$ is a log likelihood in the sense that $E(\exp(l_D(\mu)) = 1$ subject to integrability conditions; in fact, it is the dual likelihood [Mykland (1995)]. In the nonlinear case, $l_D(\mu)$ is typically not a log likelihood. For functions of means, however, DiCiccio, Hall and Romano (1991) show Bartlett correctability of R^2, so property (2.4) holds, at least, for $q = 3$ and 4.

This does not remain the case, however, for higher values of q. To see this, consider the simple case of i.i.d. bivariate data (X_i, Y_i), with a function-of-means null hypothesis on the form H_0: $E(X) - f(E(Y)) = 0$. One needs to look at bivariate data because in the scalar case, $f(E(X)) = \theta$ can be rewritten $E(X) = f^{-1}(\theta)$, which is, indeed, linear in $E(X)$.

A straightforward calculation shows that

$$k_5 = -15 f''(E(Y))\,\mathrm{cum}(Z, Z, Y) + \cdots, \tag{4.2}$$

where $Z = X - f'(E(Y))Y$ and where the expansion is in $f''(E(Y))^2$, $f'''(E(Y))$, and so on. Using (3.4), it follows that $\mathrm{cum}_5(R) = n^{-3/2}\kappa_5 + O(n^{-5/2})$, where

$$\kappa_5 = -3 f''(E(Y))\,\mathrm{cum}(Z, Z, Y)/\operatorname{Var}(Z)^{5/2} + \cdots. \tag{4.3}$$

Hence, already for $q = 5$, the R-statistic from empirical likelihood ceases, in the general case, to behave as if it were derived from a true likelihood.

4.2. *Point process likelihoods.* Superficially, the picture for nonparametric point process likelihoods [see, e.g., Andersen, Borgan, Gill and Keiding (1993)] is very similar. For problems that are linear in the cumulative hazard, the dual criterion function is a likelihood [Mykland (1995)], but this is not so in more complicated cases. Consider, by analogy to the above, the problem of comparing two survival distributions at a specific point in time. Suppose the cumulative hazards of two populations are given by $\Lambda_1(t)$ and $\Lambda_2(t)$, that numbers at risk are, respectively, $Y_1(t)$ and $Y_2(t)$, and let failures be denoted by $S_1, S_2, \dots$ and $T_1, T_2, \dots$. It is easy to see [Andersen, Borgan, Gill and Keiding (1993, Section II.7)] that the nonparametric log likelihood is given by

$$\begin{aligned} l_N(\Lambda_1, \Lambda_2) = {} & \sum_{S_i \le t} \log \Lambda_1\{S_i\} + \sum_{T_i \le t} \log \Lambda_2\{T_i\} \\ & - \int_0^t Y_1(s)\, d\Lambda_1(s) - \int_0^t Y_2(s)\, d\Lambda_2(s) + C, \end{aligned} \tag{4.4}$$

where C is random but a function of the data only. Note that one arrives at (4.4) by first setting up the likelihood assuming that the Λ's are continuous. One then shows that the maximizing values of Λ_1 and Λ_2 must have their mass concentrated on the failure times. This is a standard procedure; the unrestricted maximum likelihood estimators are the Nelson–Aalen estimators for the two populations.

The natural null hypothesis is $\Lambda_1(t) = f(\Lambda_2(t))$ (where we are holding t fixed). For example, suppose $\overline{F}_i(t)$ is the probability of surviving beyond time t, and one wishes a confidence interval for the difference in this survival probability. The null hypothesis would then be $\overline{F}_1(t) = \overline{F}_2(t) + \delta$. Assuming continuity of the cumulative hazards, $\overline{F}_i(t) = \exp(-\Lambda_i(t))$, and so this hypothesis is on the above form with $f(\Lambda) = -\log(\exp(-\Lambda) + \delta)$.

The nonparametric likelihood ratio statistic is then given by

$$\tfrac{1}{2}R^2 = \max l_N(\Lambda_1, \Lambda_2) - \max_{\Lambda_1(t) = f(\Lambda_2(t))} l_N(\Lambda_1, \Lambda_2). \tag{4.5}$$

To actually find R^2, one would maximize

$$l_L(\Lambda_1, \Lambda_2, \mu) = l_N(\Lambda_1, \Lambda_2) - n\mu(\Lambda_1(t) - f(\Lambda_2(t))) \tag{4.6}$$

in Λ_1 and Λ_2, and then let μ be such that the constraint $\Lambda_1(t) = f(\Lambda_2(t))$ be satisfied. In the spirit of dual likelihood, however, we maximize instead (4.6) subject to any μ. This gives rise (by simple differentiation) to maximizers satisfying $\hat{\Lambda}_{1,\mu}\{S_i\} = (Y_1(S_i)+n\mu)^{-1}$ and $\hat{\Lambda}_{2,\mu}\{T_i\} = (Y_2(T_i)-n\mu f'(\hat{\Lambda}_{2,\mu}(t))^{-1}$. The value $\mu = 0$ corresponds to unrestricted maximization.

Substituting $\hat{\Lambda}_{1,\mu}$ and $\hat{\Lambda}_{2,\mu}$ into (4.6) gives

$$l_L(\hat{\Lambda}_{1,0}\hat{\Lambda}_{2,0}, 0) - l_L(\hat{\Lambda}_{1,\mu}, \hat{\Lambda}_{2,\mu}, \mu) = l_D(\mu), \tag{4.7}$$

where $l_D(\mu)$ is the dual criterion function

$$\begin{aligned} l_D(\mu) = \sum_{S_i \le t} \log\left(1 + \frac{n\mu}{Y_1(S_i)}\right) + \sum_{T_i \le t} \log\left(1 - \frac{n\mu f'(\hat{\Lambda}_{2,\mu}(t))}{Y_2(T_i)}\right) \\ - n\mu g(\hat{\Lambda}_{2,\mu}(t)), \end{aligned} \tag{4.8}$$

with $g(\Lambda) = f(\Lambda) - f'(\Lambda)\Lambda$.

Obviously, $l_D(\mu)$ is maximized when $\hat{\mu}$ is such that the constraint $\hat{\Lambda}_{1,\mu}(t) = f(\hat{\Lambda}_{2,\mu}(t))$ is satisfied, and hence the R^2 from (4.5) is the same as $2(l_D(\hat{\mu}) - l_D(0))$.

An approximating log likelihood is given by

$$\begin{aligned} \tilde{l}_D(\mu, \theta) = \sum_{S_i \le t} \log\left(1 + \frac{n\mu}{Y_1(S_i)}\right) \\ + \sum_{T_i \le t} \log\left(1 - \frac{n\theta}{Y_2(T_i)}\right) - n\mu\Lambda_1(t) + n\theta\Lambda_2(t). \end{aligned} \tag{4.9}$$

Then

$$\begin{aligned} l_D(\mu) &= \tilde{l}_D(\mu, \mu f'(\hat{\Lambda}_{2,\mu}(t))) \\ &\quad + n\mu\{f(\Lambda_2(t)) - f(\hat{\Lambda}_{2,\mu}(t)) + f'(\hat{\Lambda}_{2,\mu}(t))(\hat{\Lambda}_{2,\mu}(t) - \Lambda_2(t))\} \\ &= \tilde{l}_D(\mu, \mu f'(\Lambda_2(t))) + O_Q(1) \end{aligned} \tag{4.10}$$

by Taylor expansion, where $\log(dQ/dP) = \tilde{l}_D(\mu, \mu f'(\Lambda_2(t)))$, since

$$\hat{\Lambda}_{2,\mu}(t) - \Lambda_2(t) = -n^{-1}(\partial\tilde{l}_D/\partial\theta)(\mu, \mu f'(\Lambda_2(t))) + O_Q(n^{-3/2}) \tag{4.11}$$

in asymptotically ergodic circumstances.

From (4.10) and the development in Section 3, it follows that the dual criterion function l_D satisfies (3.2) for all q, and hence the R statistic behaves, to the order under study, like one from a parametric likelihood.

APPENDIX

A.1. *Proof of Theorem* 1. (ii) implies (iii) by using Theorem 3.2.1, page 67, of Jensen (1995). Then (iii) implies (i) because, otherwise, there is a q, $3 \leq q < p$, so that $\kappa_q \neq 0$. This gives two different expansions on the form (2.6) (for p and q), and both cannot hold.

Now (i) implies (ii) because of the following. Suppose r is of order $O(n^{1/2-1/p})$ or smaller. Let $s = r/\sqrt{n}$, so $s = O(n^{-1/p})$ or smaller. Also let $\tau_n = \hat{\tau}_n/\sqrt{n}$. By definition,

$$s = \sum_{q=1}^{p} \frac{1}{(q-1)!}(\mathrm{cum}_q(T_n)n^{(q-2)/2})\tau_n^{q-1} + \frac{1}{p!}\psi_n^{(p+1)}(\tau_n^*)\tau_n^p, \tag{A1.1}$$

where $\tau_n^* \varepsilon \operatorname{int}(0, \tau_n)$. It follows that $s = \tau_n + \kappa_p \tau_n^{p-1}/(p-1)! + O(\tau_n^p) + O(n^{-1})$, and so $\tau_n = s - \kappa_p s^{p-1}/(p-1)! + O(n^{-1})$, $\tau_n^2 = s^2 - 2\kappa_p s^p/(p-1)! + O(n^{-1}s) + O(n^{-2})$, and $\tau_n^p = s^p + O(n^{-1}s)$. Hence

$$K_n''(\hat{\tau}_n) = 1 + O(s^{p-1}) \tag{A1.2}$$

and

$$\begin{aligned} K_n(\hat{\tau}_n) - \hat{\tau}_n K_n'(\hat{\tau}_n) &= -\frac{1}{2}\tau_n^2 n - \frac{p-1}{p!}\kappa_p \tau_n^p n + O(\tau_n^2) + O(\tau_n^{p+1} n) \\ &= -\frac{1}{2}s^2 n + \frac{1}{p!}\kappa_p s^p n + O(s) + O(n^{-1}) \end{aligned} \tag{A1.3}$$

from which the result follows.

A.2. *Derivation of the likelihood results in Section* 3. Consider first why (3.2) ($q < p$) implies (2.4). Set $l(\theta) = \tilde{l}(\theta) + d(\theta)$, where

$$\exp\bigl(d(\theta)\bigr) = E_{\theta_0} \exp\bigl(l(\theta) - l(\theta_0)\bigr). \tag{A2.1}$$

Hence, $E_{\theta_0} \exp(\tilde{l}(\theta) - \tilde{l}(\theta_0)) = 1$, and so $\tilde{l}(\theta)$ satisfies the Bartlett identities exactly at the value θ_0 [cf. Chapter 7.2 of McCullagh (1987)].

Consider now the stochastic expansions for R and $\tilde{R}$. Let $Z^{(q)} = (\tilde{l}^{(q)} - E(\tilde{l}^{(q)}))/\mathrm{sd}(\tilde{l}^{(q)})$, evaluated at $\theta = \theta_0$, and suppose that these standardized sums are asymptotically normal for $q \leq p$. By the same development as on page 214 of McCullagh (1987), but to suitable order, $\tilde{R}$ has the expansion

$$\tilde{R} = \tilde{Q}^{(0)} + n^{-1/2}\tilde{Q}^{(1)} + \cdots. \tag{A2.2}$$

Here,

$$\begin{aligned} \tilde{Q}^{(k)} = {} & \text{const } \times \; Z_1 \cdots Z_{k+1} \\ & + \text{a linear combination of products } Z_1 \cdots Z_j, \end{aligned} \tag{A2.3}$$

where $1 \le j \le k$ and the Z_i are chosen from $Z^{(1)}, \ldots, Z^{(k+1)}$. Similarly, R has a stochastic expansion

$$R = Q^{(0)} + n^{-1/2}Q^{(1)} + \cdots. \tag{A2.4}$$

By (3.2), however, the leading term in $\widetilde{Q}^{(k)}$, that is, const × $Z_1 \cdots Z_{k+1}$, is the same as for $Q^{(k)}$. Hence, the expressions for $\operatorname{cum}_q(R)$ and for $\operatorname{cum}_q(\widetilde{R})$ only differ through terms of the form (const ×)

$$\begin{aligned}
&\operatorname{cum}\bigl(n^{-k_1/2} Z_1^{(1)} \cdots Z_{r_1}^{(1)}, \ldots, n^{-k_q/2} Z_1^{(q)}, \ldots, Z_{r_q}^{(q)}\bigr) \\
&\quad = n^{-1/2\sum k_i} \operatorname{cum}\bigl(Z_1^{(1)} \cdots Z_{r_1}^{(1)}, \ldots, Z_1^{(q)} \cdots Z_{r_q}^{(q)}\bigr) \\
&\quad = O\bigl(n^{-1/2\sum k_i + 1/2\sum r_i - q + 1}\bigr) \\
&\quad = O\bigl(n^{-(q-1)/2}\bigr),
\end{aligned} \tag{A2.5}$$

since $1 \le r_i \le k_i + 1$ and since there is at least one i for which $1 \le r_i \le k_i$. Hence $\operatorname{cum}_q(R) = O(n^{-(q-1)/2})$. Then $\operatorname{cum}_q(R) = O(n^{-q/2})$ follows if (2.1) is valid.

To see why (3.3) must now lead to (3.4), write

$$l(\theta) = \tilde{l}(\theta) + \frac{1}{p!} n k_p (\theta - \theta_0)^p + O_p\bigl(n(\theta - \theta_0)^{p+1}\bigr),$$

where $\tilde{l}(\theta)$ satisfies (3.2) at θ_0 for $q \le p$. If $\hat{\theta}$ and $\tilde{\theta}$ are the maximum likelihood estimators for $l(\theta)$ and $\tilde{l}(\theta)$, respectively, then it is easy to see that $\hat{\theta} - \tilde{\theta} = O_p(n^{-(p-1)/2})$. It follows that, since $p \ge 3$,

$$\begin{aligned}
\frac{1}{2}R^2 &= \frac{1}{2}\widetilde{R}^2 + \frac{1}{p!} n k_p (\hat{\theta} - \theta_0)^p + O_p\bigl(n^{-(p-1)/2}\bigr) \\
&= \frac{1}{2}\widetilde{R}^2 + \frac{1}{p!} k_p \sigma^{-p} n^{-(p-2)/2} \widetilde{R}^p + O_p\bigl(n^{-(p-1)/2}\bigr).
\end{aligned}$$

Hence,

$$R = \widetilde{R} + \frac{1}{p!} k_p \sigma^{-p} n^{-(p-2)/2} \widetilde{R}^{p-1} + O_p\bigl(n^{-(p-1)/2}\bigr),$$

and so

$$\begin{aligned}
\operatorname{cum}_p(R) &= \operatorname{cum}_p(\widetilde{R}) + \frac{1}{p!} k_p \sigma^{-p} n^{-(p-2)/2} \operatorname{cum}\bigl(\widetilde{R}^{p-1}, \widetilde{R}, \ldots, \widetilde{R}\bigr) + O_p\bigl(n^{-p/2}\bigr) \\
&= \frac{1}{p} k_p \sigma^{-p} n^{-(p-2)/2} + O_p\bigl(n^{-p/2}\bigr).
\end{aligned}$$

Hence the result follows.

Acknowledgments. The author thanks Ole Barndorff-Nielsen, Anthony Davison, Tom DiCiccio, Jens Jensen, Donald Pierce, Ib Skovgaard, Trevor Sweeting, Andy Wood and the referee and Associate Editor for extremely useful and at times highly entertaining discussions. I am also very grateful for my discussions with Stephen Corcoran, who has since so tragically passed away—he will be a loss to our field! I also thank Mitzi Nakatsuka for typing the paper.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models based on Counting Processes.* Springer, New York.

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

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. and COX, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *J. Roy. Statist. Soc. B* **41** 279–312.

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

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.

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

CORCORAN, S. A., DAVISON, A. C. and SPADY, R. H. (1995). Reliable inference from empirical likelihoods. Preprint.

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

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B* **49** 1–18.

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

DAVISON, A. C., HINKLEY, D. V. and WORTON, B. J. (1992). Bootstrap likelihoods. *Biometrika* **79** 113–130.

DICICCIO, T. J., HALL, P. and ROMANO, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19** 1053–1061.

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.

GODAMBE, V. P. and HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Int. Statist. Rev.* **55** 231–244.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

JACOD, J. (1975). Multivariate point processes: predictable projection, Radon-Nikodym derivatives, representation of martingales. *Z. Wahrsch. Verw. Gebiete* **31** 235–253.

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 University Press.

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

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

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. (1999). Empirical likelihood in the presence of nuisance parameters. *Biometrika* **86** 203–211.

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

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

MCCULLAGH, P. and TIBSHIRANI, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. 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. (1995). Dual likelihood. *Ann. Statist.* **23** 396–421.

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

ROBINSON, J., HÖGLUND, T., HOLST, L. and QUINE, M. P. (1990). On approximating probabilities for small and large deviations on R^d. *Ann. Probab.* **18** 727–753.

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

SKOVGAARD, I. (1991). *Analytic Statistical Models.* IMS, Hayward, CA.

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

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

WALLACE, D. L. (1958). Asymptotic approximations to distributions. *Ann. Math. Statist.* **29** 635–654.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447.

WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123.

WONG, W. H. and SEVERINI, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Ann. Statist.* **19** 603–632.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
E-MAIL: mykland@galton.uchicago.edu