Bernoulli 2(2), 1996, 145–165

An explicit large-deviation approximation to one-parameter tests

IB M. SKOVGAARD

Department of Mathematics and Physics, The Royal Veterinary and Agricultural University, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark

An approximation is derived for tests of one-dimensional hypotheses in a general regular parametric model, possibly with nuisance parameters. The test statistic is most conveniently represented as a modified log-likelihood ratio statistic, just as the R^*-statistic from Barndorff-Nielsen (1986). In fact, the statistic is identical to a version of R^*, except that a certain approximation is used for the sample space derivatives required for the calculation of R^*. With this approximation the relative error for large-deviation tail probabilities still tends uniformly to zero for curved exponential models. The rate may, however, be $O(n^{-1/2})$ rather than $O(n^{-1})$ as for R^*. For general regular models asymptotic properties are less clear but still good compared to other general methods. The expression for the statistic is quite explicit, involving only likelihood quantities of a complexity comparable to an information matrix. A numerical example confirms the highly accurate tail probabilities. A sketch of the proof is given. This includes large parts which, despite technical differences, may be considered an overview of Barndorff-Nielsen's derivation of the formulae for p^* and R^*.

Keywords: conditional inference, large-deviation expansions, modified log-likelihood ratio test, nuisance parameters, parametric inference

1. Introduction

The purpose of the present paper is to derive an explicit general approximation for testing a one-dimensional, possibly composite, hypothesis in a well-behaved parametric model. By a one-dimensional hypothesis is meant a hypothesis that a single coordinate of the parameter vector assumes a particular value.

The entire approach and the result is highly related to and based on the line of theory developed by Barndorff-Nielsen through the p^* and the R^* formulae (see, in particular, Barndorff-Nielsen 1980; 1986; 1991). Thus, the paper deals with likelihood inference, and what is described may be seen as an attempt to improve some of the classical, normal-based, asymptotic results, especially the chi-squared approximation to minus twice the log-likelihood ratio statistic. In the case of a one-dimensional hypothesis the signed square root of this statistic has a standard normal distribution under the hypothesis, and the simplest way of presenting the present result is as a modification of this statistic, quite analogously to R^*. This statistic, here denoted $\tilde{R}$, is given in formula (2) in combination with formula (1) in Section 3.

The standard normal approximation to the tail probabilities of the distribution of $\tilde{R}$ is of

1350-7265

the large-deviation type for curved exponential models, i.e., the relative error of the tail probability tends to zero uniformly in a region of large deviations. The rate at which this relative error tends to zero is at least as good as $O(n^{-1/2})$ under repeated sampling. For comparison, the relative error for R^* is $O(n^{-1})$.

The calculation of $\tilde{R}$ involves only likelihood quantities of a computational complexity comparable to that of the Fisher information matrix. These quantities are well defined for any well-behaved parametric model, and the result is therefore not confined to curved exponential families. However, the proof is presented for curved exponential families, and for more general models some asymptotic results then follow by approximation to curved exponential families (see Section 4).

The reason why the R^* formula cannot be applied to the problem addressed here is that this statistic involves some sample space derivatives (see Barndorff-Nielsen 1991, Sections 2 and 3). These are derivatives of likelihood quantities with respect to the maximum likelihood estimate, through the minimal sufficient statistic. When the maximum likelihood estimate is not sufficient an ancillary statistic must be specified for the calculation of such sample space derivatives. The non-uniqueness of ancillaries that may be used for this purpose and the difficulties involved in their specification make it difficult to calculate R^* in general.

Barndorff-Nielsen and Chamberlin (1991; 1994) solve the problem by approximating R^*, but at the price of losing the general large-deviation properties of the approximation. The same is true for DiCiccio and Martin (1993), although their method of approximation is quite different.

Jensen (1992) constructs a statistic similar to R^* for curved exponential families by specifying the ancillary statistic as a series of one-dimensional signed log-likelihood ratios from the full exponential model down to the curved model. By means of this construction he can prove remarkably good large-deviation properties of the approximation, but the result may be hard to compute and may depend on the series of one-dimensional hypotheses chosen.

The main idea behind the present approach is that an explicit general approximation can be made to the sample space derivatives required. This approximation is sufficiently good to keep large-deviation properties and sufficiently simple to make it possible to derive explicit results.

In fact, in the case of no nuisance parameters, the approximate sample space derivatives used here are identical to the sample space derivatives in Fraser and Reid (1988), who define the ancillary statistic in terms of the sample space derivatives of the log-likelihood. The present use of these derivatives is different though, namely as approximations to sample space derivatives arising from ancillaries that are directed versions of the log-likelihood ratio for testing the model against the full exponential family.

After the introduction of notation in Section 2, the main result is given in Section 3. Its properties, especially in terms of asymptotic errors, are described in Section 4. This part of the paper, possibly together with the numerical example in Section 5, should suffice for readers who are not interested in the methods and proofs used. The proof is outlined for curved exponential families, for which notation and basic concepts are introduced in Section 6. A conceptual and mathematical description of ancillaries and sample space derivatives is given in Section 7, while the crucial approximation to the sample space derivatives is discussed in Section 8. With this approximation substituted into the expression for R^* in Barndorff-Nielsen (1991) the result might be derived, but this would not reveal the accuracy of the result. Partly for this reason, an entire outline of the proof from the beginning is given in Section 9. Another reason for this is that some results in this section may be of independent interest, in particular a number of intermediate results that hold for any ancillary statistic, regardless of its distributional properties. Finally, the discussion in Section 10 mainly points out some further problems.

2. Set-up and notation

Let $\{f(y;\beta);\beta\in B\subseteq\mathbb{R}^p\}$ be a family of densities of the random variable Y, indexed by the p-dimensional parameter β, possibly restricted to some subset B. The domain of Y and the underlying measure are of no importance in the present connection, except that the asymptotic results require absolute continuity of the distribution of the canonical sufficient statistic with respect to Lebesgue measure.

The problem is to derive a test for a one-dimensional hypothesis. More specifically, let

$$\beta=(\beta_1,\ldots,\beta_{p-1},\beta_p)=(\alpha,\psi),$$

where $\alpha=(\beta_1,\ldots,\beta_{p-1})$ and $\psi=\beta_p$; we wish to test the hypothesis

$$H_0:\psi=\psi_0.$$

Let $\hat{\beta}(y)=\hat{\beta}=(\hat{\alpha},\hat{\psi})$ denote the maximum likelihood estimate of the full parameter vector, and let $\tilde{\beta}(y)=\tilde{\beta}=(\tilde{\alpha},\psi_0)$ denote the maximum likelihood estimate under the hypothesis.

The log-likelihood function is denoted

$$\ell(\beta)=\ell(\beta;y)=\log f(y;\beta),$$

where the first version is used when y or the value of some sufficient statistic is understood. The kth derivative of the log-likelihood function is denoted D_k, i.e.,

$$D_k(\beta)=D_k(\beta;y)=\frac{\partial^k}{\partial\beta^k}\,\ell(\beta;y),$$

which is a k-sided array with p^k entries. In particular, D_1 is the score function.

The cumulants of the log-likelihood derivatives are denoted by χ's, with indices corresponding to the derivatives in question. Thus, for example,

$$\chi_k(\beta)=E_\beta\{D_k(\beta;Y)\},\qquad \chi_{km}(\beta)=\operatorname{cov}_\beta\{D_k(\beta;Y),D_m(\beta;Y)\},$$

denote means and variances. For the more common information quantities we use the special notation

$$i(\beta)=\chi_{11}(\beta)=-\chi_2(\beta),\qquad j(\beta;y)=-D_2(\beta;y).$$

In particular, $i(\beta)$ is the (expected) Fisher information.

We use abbreviations such as $\hat{\ell} = \ell(\hat{\beta})$, $\hat{i} = i(\hat{\beta})$, $\hat{j} = j(\hat{\beta}; y)$, and $\tilde{j} = j(\tilde{\beta}; y)$, and frequently omit the argument y. Note that $\hat{j}$ is the observed Fisher information.

Finally, we need two less familiar quantities, $\hat{q}$ and $\hat{S}$, defined below. These are based on covariances of likelihood differences and derivatives. More specifically, let

$$\chi_{10}(\beta_1, \beta_2; \beta) = \mathrm{cov}_\beta\{D_1(\beta_1; Y), \ell(\beta_1; Y) - \ell(\beta_2; Y)\},$$

and

$$\chi_{11}(\beta_1, \beta_2; \beta) = \mathrm{cov}_\beta\{D_1(\beta_1; Y), D_1(\beta_2; Y)\},$$

and define

$$\hat{q} = \chi_{10}(\hat{\beta}, \tilde{\beta}; \hat{\beta}), \qquad \hat{S} = \chi_{11}(\hat{\beta}, \tilde{\beta}; \hat{\beta}).$$

Note that $\hat{q}$ is a p-vector, while $\hat{i}$, $\hat{j}$ and $\hat{S}$ are $p \times p$ matrices.

The determinant of a matrix, M say, is denoted $|M|$, and its transpose is denoted M^{T}.

3. The approximate test

Standard asymptotic theory would suggest an asymptotic chi-squared test based on twice the log-likelihood ratio

$$R^2 = 2\{\ell(\hat{\beta}) - \ell(\tilde{\beta})\}.$$

Alternatively, one-sided tests may be calculated from the standard normal distribution of the directed log-likelihood ratio test statistic R, equipped with the sign of $\hat{\psi} - \psi_0$.

Far better asymptotic performance is obtained by use of the modified signed log-likelihood ratio introduced by Barndorff-Nielsen (1986; 1991) and given by

$$R^* = R - \frac{1}{R}\log(R/U),$$

where U is a quantity which unfortunately is difficult to calculate since it requires specification of an ancillary statistic – or at least a local specification of the change of the log-likelihood difference with a change of the estimate. When the estimate is not sufficient this 'sample space derivative' is only defined in the conditional distribution given a supplementary statistic which together with the estimate is sufficient. When this supplementary statistic is chosen in a certain way the standard normal approximation to R^* becomes accurate to third order, i.e., with an error of order $O(n^{-3/2})$ in repeated sampling of n observations in a well-behaved model (see Barndorff-Nielsen 1986; 1991).

The point of the present paper is to provide an approximation, $\tilde{U}$, to U, which is easily calculated and sufficiently accurate to maintain the high-quality asymptotic behaviour, although one order of magnitude may be lost compared to R^*. This approximation is

$$\tilde{U} = [\hat{S}^{-1}\hat{q}]_p |\hat{j}|^{1/2} |\tilde{j}_{\alpha\alpha}|^{-1/2} |\hat{i}|^{-1} |\hat{S}|, \tag{1}$$

where $[\cdots]_p$ denotes the pth coordinate of the vector, and $\tilde{j}_{\alpha\alpha}$ denotes the upper left $(p-1) \times (p-1)$ submatrix of $\tilde{j}$. Notice that $\tilde{j}_{\alpha\alpha}$ is simply the observed Fisher information for the parameter α under the hypothesis.

Insertion in the expression for R^* now defines the statistic

$$\tilde{R} = R - \frac{1}{R}\log(R/\tilde{U}) \tag{2}$$

and the claim is, as for R^*, that a standard normal distribution provides an accurate approximation to its distribution under the hypothesis. This statement will be made more precise in the following section.

It may be noted that in terms of asymptotic approximation an entirely equivalent result may be obtained by use of a different type of Laplace approximation to a tail integral, using a method from Bleistein (1966), also known from Lugannani and Rice (1980). This gives the right tail probability as

$$1 - \Phi(R) + \frac{\phi(R)}{R}(R/\tilde{U} - 1), \tag{3}$$

instead of

$$1 - \Phi(\tilde{R}).$$

The asymptotic equivalence of the two expressions is proved in Jensen (1992, Lemma 2.1). Numerical examples seem to indicate, however, that the R^* version is preferable (see, for example, Pierce and Peters 1992).

The quantities $\hat{q}$ and $\hat{S}$ are usually of the same computational complexity as $\hat{\imath}$, since they are also covariances of ordinary likelihood quantities. Alternative expressions for $\hat{q}$ and $\hat{S}$ are in terms of derivatives of the Kullback–Leibler distance

$$KL(\beta, \beta_1) = E_\beta\{\log f(Y;\beta) - \log f(Y;\beta_1)\},$$

from which we obtain

$$\chi_{10}(\beta, \beta_1; \beta) = \frac{\partial}{\partial\beta} KL(\beta; \beta_1)$$

and

$$\chi_{11}(\beta, \beta_1; \beta) = -\frac{\partial}{\partial\beta}\frac{\partial}{\partial\beta_1} KL(\beta; \beta_1).$$

The first of these derivatives also appears in Sweeting (1995, Section 5).

4. Properties of the approximation

Since the expression for $\tilde{R}$ only involves likelihood quantities it is well defined for all sufficiently regular parametric models and does not require an embedding in an exponential family. Furthermore, it is trivially invariant under sufficient transformations of the data.

The expression is also invariant under relevant smooth one-to-one transformations of the parameter, i.e., under transformations of the parameter of interest, ψ, and under transformations of β preserving ψ. There is no loss of generality in the formulation of the hypothesis as a hypothesis concerning a single coordinate since we can always make the one-dimensional parameter of interest a coordinate.

Concerning the asymptotic properties, we confine ourselves to the case of n independent replications, although the approximation may well also be reasonable in other cases. The results are stated in terms of right tail probabilities, but hold analogously for the left tail. The asymptotic results are only valid for absolutely continuous distributions.

Assume first that the model is a 'curved exponential family', i.e., a smooth submodel of an exponential family. Let $\bar{t}$ denote the mean of the canonical sufficient statistic and let $\bar{W} = \bar{W}(\bar{t})$ denote n^{-1} times minus twice the log-likelihood ratio test statistic for testing the model against the full exponential family. Scaling by the divisor n ensures that $\bar{W}$ depends only on $\bar{t}$, not on n. The conditional result below holds given any ancillary statistic of the form $A = A(\bar{t})$ which is such that $\bar{W}$ is a function of A, combined with the requirement that $(\hat{\beta}, A)$ is sufficient. Thus A is any 'directed log-likelihood ratio' as opposed to the ancillaries used for R^* which are directed *modified* log-likelihood ratios. Any statistic A of the kind used in the present paper will generally be a first-order ancillary statistic in the sense that its standardized distribution is free of β under the model apart from a term of order $n^{-1/2}$.

Now the result for the conditional tail probability, given any of the ancillary statistics of the type mentioned above, is that, under the hypothesis,

$$1 - \Phi(\tilde{r}) = \mathrm{pr}_\beta\{\tilde{R} \geq \tilde{r} \,|\, A\}\{1 + O(n^{-1}) + O(\|\hat{\psi} - \psi_0\| \, \|A\|)\}$$

as $n \to \infty$ uniformly over $(\hat{\psi}, A)$ in some fixed neighbourhood of $(\psi_0, 0)$, where $A = 0$ is chosen to correspond to $\bar{W} = 0$. Since $(\hat{\psi}, A)$ converges to its mean at rate $n^{-1/2}$, this is a large-deviation region. Notice, however, that both of the normal deviations of A and $\hat{\psi} - \psi_0$ are of order $n^{-1/2}$, but either or both of them may become of order $O(1)$ in a large-deviation region. Thus the result states that the error is of order n^{-1} in a normal-deviation region, whereas the relative error is of order $n^{-1/2}$ in a large-deviation region of either the ancillary or the estimate, but not both.

Unconditionally this implies that the relative error is $O(n^{-1/2})$ uniformly in a fixed set of values of $\hat{\psi}$, i.e., in a large-deviation region. For comparison the relative error for R^* is $O(n^{-1})$. For the numerical quality of the approximation it is presumably more important, however, that the large-deviation property is kept, since this results in a far better tail behaviour of the approximation than the central type of expansions that form the basis of standard asymptotic theory.

For models that are not submodels of exponential models it is more difficult to derive conditional results since it is hard to come up with useful ancillary statistics in a general form. However, for analytic models (see Skovgaard 1990), it follows that unconditionally we have

$$1 - \Phi(\tilde{r}) = \mathrm{pr}_\beta\{\tilde{R} \geq \tilde{r}\}\left\{1 + O(n^{-1} + n^{-1/2}\|\hat{\psi} - \psi_0\| + n^{-K}\mathrm{pr}_\beta\{\tilde{R} \geq \tilde{r}\}^{-\epsilon})\right\}$$

for any $\epsilon > 0$ and $K > 0$, uniformly in a fixed set of $\hat{\psi}$ around ψ_0. With $\epsilon = 0$ the uniform relative error would have been kept, so the result is that this is almost the case. Since any analytic model can be approximated locally to any order by a curved exponential family, this result follows almost immediately from this general approximation theory in Skovgaard (1990, Section 2.7), but the proof will not be given here. That the normal deviation error is of order $O(n^{-1})$ is a trivial consequence, because then $\|\hat{\psi} - \psi_0\| = O(n^{-1/2})$.

As a final property it is worth noting that $\tilde{R}$ agrees exactly with R^* when no ancillary variable is necessary, which essentially is when the model is a full exponential family. However, the hypothesis may be curved, so for the Behrens–Fisher example given in Jensen (1992) the two statistics are identical.

5. An example

As a numerical example we consider a one-way analysis of variance with random effects. Thus, let $i = 1, \dots, m$ enumerate the groups and $j = 1, \dots, n_i$ the observations within groups, and let X_{ij} be normally distributed with mean μ and variance ω^2. Observations from different groups are assumed to be independent, while the within-group correlation is ρ. We allow for negative correlations such that the lower bound for ρ becomes

$$\rho > -(n_{\max} - 1)^{-1},$$

where $n_{\max}$ is the largest group size. Notice that the variance of the largest group mean tends to zero as ρ approaches its lower bound.

Usually this random effects model is written in the form

$$X_{ij} = \mu + B_i + \epsilon_{ij}, \tag{4}$$

where the B_i's and the ϵ_{ij}'s are all independent with standard deviations σ and σ_B, respectively. This formulation covers only cases with non-negative within-group correlations, however.

We wish to test the hypothesis $\mu = 0$ and consider three test statistics. First, R denotes the unmodified directed square root of the log-likelihood ratio statistic. Second, the modified statistic $\tilde{R}$ from (2) is considered. Third, we compute an approximate F-test statistic based on the approximation suggested by Satterthwaite (1946). This test is based on the ratio of the between-group mean square to its estimated expectation which is a linear combination of the same mean square and the residual mean square. The distribution of the ratio is then approximated by an F-distribution for which the number of degrees of freedom for the denominator is chosen to make the variance of the denominator correct. This third statistic, denoted 'Satterthwaite' in Table 1, may not be as attractive as the likelihood ratio based statistics for variance components in general, but it has the advantage of admitting Satterthwaite's approximation to the null distribution – an approximation known to be quite accurate.

The computation of $\tilde{R}$ is straightforward and uninteresting, so it will not be shown here. It involves nothing beyond matrix inversions of the size of the information matrix, i.e. 3 × 3 matrices, and calculation of cumulants of order 4 or less in a normal distribution.

The distributions are, like the entire model, symmetric about zero, so only two-sided tail probabilities will be considered. The approximate F-statistic is two-sided by construction although it might easily be reformulated as a one-sided t-test.

Only one set of simulations will be shown. Others have been done which indicate the same accuracy of the approximations. The example has 5 groups of sizes 1, 2, 3, 4, and 5,

Table 1. Numbers of exceedances of nominal two-sided significance levels in 100 000 simulations of the one-way analysis of variance with random effects; the nominal significance levels are given indirectly as the 'expected' numbers of exceedances

Expected	50 000	20 000	10 000	2000	200	20	2
R	56 635	27 771	16 399	4780	917	146	27
$\tilde{R}$	50 061	20 229	10 128	2101	224	24	3
Satterthwaite	50 470	19 948	9875	1904	181	13	1

respectively, and 100 000 samples were generated from the model (4) with $\mu = 0$, $\sigma = 1.0$, and $\sigma_B = 0.04$. The pseudo-random number generator RAN2 from Press *et al.* (1986, Chapter 7) was used. The numbers of exceedances of nominal two-sided tail probabilities are seen in Table 1, together with the nominal expected numbers according to the respective approximate distribution.

The usual asymptotic approximation to the uncorrected log-likelihood ratio statistic, R, rejects the hypothesis far too often in the tails. For the assessment of $\tilde{R}$ this shows that although the example is a fairly 'nice' one, standard likelihood asymptotic approximations are not automatically of high quality. In contrast, the adjusted log-likelihood ratio test, $\tilde{R}$, behaves well, even in the extreme tail. This type of behaviour is typical of approximations with large-deviation properties as opposed to central approximations such as for R.

The approximation based on Satterthwaite's method is of the same quality as $\tilde{R}$. This method is known to give quite accurate approximations, but is limited to the approximate F-tests, which may not be the most desirable tests in variance component models in general. One reason is that there are many ways of constructing such tests for more complicated models.

There was one case, not counted in the table, for which the two likelihood methods broke down because the numerical algorithm searching for the solution to the likelihood equation did not converge. In all other cases such a solution was found. However, strictly speaking, the likelihood methods were inapplicable in all cases, because the likelihood function tends to infinity when μ is held equal to the mean of the largest group while the within-group correlation tends to its lower bound. This is just one reason why the method of maximum likelihood should not be used for variance component models; the more structured approach of restricted maximum likelihood should be used instead. It is an important challenge to extend the type of methods described in the present paper to such structured inference.

6. Curved exponential families

Consider an exponential family with densities

$$f(y;\theta) = \exp\{\langle\theta, t\rangle - \kappa(\theta)\}$$

with respect to some underlying measure, where the parameter vector θ and the canonical sufficient statistic $t = t(y)$ both belong to $\mathbb{R}^k$, and $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{R}^k$.

We use the special notation

$$\tau(\theta) = \mathrm{E}_\theta\{t\}, \qquad \Sigma(\theta) = \mathrm{var}_\theta\{t\}$$

for the first two derivatives of the cumulant generating function κ. The representation is assumed to be minimal, such that $\Sigma(\theta)$ has full rank.

A curved exponential model is given by

$$\theta = \theta(\beta)$$

where $\beta \in \mathbb{R}^p$. The derivatives of the mapping θ at β are denoted $D\theta(\beta)$, $D^2\theta(\beta)$, and so on, and $D\theta$ is assumed to have rank p.

Assume that n independent observations are obtained from a distribution within this model. The mean, $\bar{t}$, of the n observations of t is sufficient, and we may then write $\bar{t}$ in place of y in all relevant statistics. The score statistic is

$$D_1(\beta;\bar{t}) = n\langle D\theta(\beta), \bar{t} - \tau(\theta(\beta))\rangle$$

where the inner product is taken over $\mathbb{R}^k$, such that D_1 becomes a p-vector. Similarly, further differentiation yields

$$D_k(\beta;\bar{t}) = n\big\{\langle D^k\theta(\beta), \bar{t} - \tau(\theta(\beta))\rangle + \chi_k(\theta(\beta))\big\}. \tag{5}$$

The important thing to notice is that all these log-likelihood derivatives deviate from their means by linear functions of the canonical sufficient statistic, $\bar{t}$.

The maximum likelihood estimate, $\hat{\beta}$, solves the likelihood equation

$$\langle D\hat{\theta}, \bar{t} - \hat{\tau}\rangle = 0$$

with obvious abbreviations.

Some of the quantities entering the expression for $\tilde{U}$ from (1) are

$$\hat{q} = n(D\hat{\theta})^{\mathrm{T}}\hat{\Sigma}(\hat{\theta} - \tilde{\theta}),$$

where $\tilde{\theta} = \theta(\tilde{\beta})$ is the estimate under the hypothesis, and

$$\hat{S} = n(D\hat{\theta})^{\mathrm{T}}\hat{\Sigma}(D\tilde{\theta}).$$

Notice that the computation of these quantities is of the same complexity as the computation of the information $n(D\theta)^{\mathrm{T}}\Sigma(D\theta)$ and essentially requires only the covariance matrix of t and the first derivative of the mapping θ.

To make it easier to see the power of the n appearing in the various expressions, we use $\bar{\ell}$ to denote the function of $\bar{t}$ that is 'free of n', i.e.,

$$\bar{\ell}(\beta;\bar{t}) = \langle\theta(\beta), \bar{t}\rangle - \kappa(\theta(\beta)) = n^{-1}\ell(\beta;\bar{t}).$$

For the same reason we define the information quantities

$$i_1(\beta) = n^{-1}i(\beta), \qquad j_1(\beta;\bar{t}) = n^{-1}j(\beta;\bar{t}),$$

which depend on β and $\bar{t}$ only, not on n.

Note that all likelihood derivatives, D_k, and their cumulants are proportional to n, when viewed as functions of β and $\bar{t}$.

There is no loss of generality in considering n independent replications since all expressions are equally valid for the special case with $n = 1$, and hence $\hat{i}_1 = \hat{i}$, and so on. Asymptotic results are, however, only proved for n independent replications.

7. Ancillaries and sample space derivatives

We continue to consider a curved exponential family, but shall from time to time, when explicitly stated, revert to more general models.

Notice first that the likelihood equation may be rewritten as the orthogonality

$$\bar{t} - \hat{\tau} \perp_{\hat{\Sigma}^{-1}} \hat{\Sigma}(D\hat{\theta}) \tag{6}$$

of the two vectors in the $\bar{t}$-space with respect to the variable inner product $\hat{\Sigma}^{-1}$. Thus the observations of $\bar{t}$ that lead to the same estimate $\hat{\beta}$ are located in a $(k-p)$-dimensional linear subspace, and the observation deviates from the estimated mean value by the vector $\bar{t} - \hat{\tau}$ in this subspace. The orthogonality, (6), of this subspace to the tangent space, spanned by $\hat{\Sigma}(D\hat{\theta})$, is the reason why it is convenient to work with the variable inner product, as will be seen in Section 9.

In the development below, which goes through the p^* formula, it is necessary to be able to write the sufficient statistic as a function of $(\hat{\beta}, A)$, where A is some supplementary statistic of dimension $k - p$. In the present paper we require that $A = A(\bar{t})$ is a smooth function of $\bar{t}$, and that $\bar{t}$ and $(\hat{\beta}, A)$ are in one-to-one correspondence. We shall refer to any such statistic as a *supplementary* statistic and reserve the word 'ancillary' for a statistic for which further properties are at least desired.

For fixed A, the sufficient statistic $\bar{t} = \bar{t}(\hat{\beta}, a)$ moves along a p-dimensional level surface of A which may be thought of as 'parallel' to the model space $\{\bar{t} = \tau(\theta(\beta))\}$ parametrized by β.

The fact that $\bar{t}$ is a function of $\hat{\beta}$ and A means that derivatives of $\bar{t}$, and consequently of likelihood quantities, may be defined with respect to $\hat{\beta}$ for fixed A. From now on we reserve a prime to denote such a derivative, i.e.,

$$t' = \frac{\partial \bar{t}}{\partial \hat{\beta}}, \qquad \ell' = \ell'(\beta; \bar{t}) = \frac{\partial}{\partial \hat{\beta}} \ell(\beta; \bar{t}(\hat{\beta}, A)).$$

These derivatives are usually referred to as *sample space derivatives*. Notice that, for example, $\hat{D}_1' = D_1'(\hat{\beta})$ means $D_1(\beta, \bar{t}(\hat{\beta}, A))$ differentiated with respect to $\hat{\beta}$ before $\hat{\beta}$ is substituted for β. Furthermore, we use the convention that a sample space derivative corresponds to the 'last index' of an array, for example to the columns of the $p \times p$ matrix D_1'. From (5) it is seen that for curved exponential families we have

$$D_k'(\beta) = n\langle D^k\theta(\beta), \bar{t}' \rangle. \tag{7}$$

As ancillary statistic, $A = A(\bar{t})$, for curved exponential families we shall consider statistics with the first or both of the following two properties:

(A1) The model subspace $\{\bar{t} = \hat{\tau}\}$ is a level surface of A. In this case we let $A = 0$ represent this subspace.

(A2) The log-likelihood ratio test statistic of the model against the full exponential family is a function of A, i.e., A is a directed log-likelihood ratio.

The asymptotic result of the paper, as stated in Section 4, requires an ancillary statistic with both of the properties.

Let $\bar{\theta}$ denote the maximum likelihood estimate of θ in the full exponential family, i.e., $\bar{\theta} = \tau^{-1}(\bar{t})$. Then minus twice the log-likelihood ratio test statistic for the model against the full exponential family is

$$W = 2n(\bar{\ell}(\bar{\theta}) - \bar{\ell}(\hat{\theta})).$$

Since the model subspace corresponds to the set $\{W = 0\}$, which is of the same dimension as any level subspace for A, Property (A2) implies Property (A1).

8. Approximation of sample space derivatives

The most important point of the present paper is to show that the sample space derivatives needed for calculation of the tail probability corresponding to the R^* formula from Barndorff-Nielsen (1986; 1991) may be approximated sufficiently well to obtain a large-deviation approximation.

To do this we first need to define a tangent space projection related to the orthogonality in (6). Thus, let $\hat{P}$ denote the orthogonal projection

$$\hat{P} = \hat{\Sigma}(D\hat{\theta})\hat{i}_1^{-1}(D\hat{\theta})^{\mathrm{T}} \tag{8}$$

on $\hat{\Sigma}(D\hat{\theta})$ with respect to the variable metric, or inner product, $\hat{\Sigma}^{-1}$. This is the projection on the subspace for $\bar{t}$ tangent to the model space at $\hat{\tau}$.

The approximation of the sample space derivative used in the present paper is now simply given by

$$\bar{t}' \approx \hat{P}(\bar{t}'), \tag{9}$$

the point being that this has the unique explicit expression

$$\hat{P}(\bar{t}') = \hat{\Sigma}(D\hat{\theta})\hat{i}_1^{-1}(D\hat{\theta})^{\mathrm{T}}\bar{t}' = \hat{\Sigma}(D\hat{\theta})\hat{i}_1^{-1}n^{-1}\hat{D}_1' = \hat{\Sigma}(D\hat{\theta})\hat{i}_1^{-1}\hat{j}_1, \tag{10}$$

since $(D\hat{\theta})^{\mathrm{T}}\bar{t}' = \langle D\hat{\theta}, \bar{t}'\rangle = n^{-1}\hat{D}_1'$. This result holds for any supplementary statistic.

For general, non-exponential, models the projection may be defined as a regression of the log-likelihood derivatives on the score statistic, i.e.,

$$P(D_k) = \chi_k + \chi_{k1}\chi_{11}^{-1}D_1, \tag{11}$$

which is then used at $\hat{\beta}$. This gives

$$\hat{P}(\hat{D}_k') = \hat{\chi}_{k1}\hat{\chi}_{11}^{-1}\hat{D}_1' = \hat{\chi}_{k1}\hat{i}^{-1}\hat{j}, \tag{12}$$

because of the well-known relation

$$\hat{D}_1' = \hat{j}$$

which is immediately obtained by differentiation of the equation $D_1(\hat{\beta}; y(\hat{\beta}, A)) = 0$ with respect to $\hat{\beta}$. This relation may be found, for example, in Barndorff-Nielsen and Cox (1994, Section 5.2).

It may easily be checked that the general definition of the projection agrees with that for curved exponential families. At the same time, it gives a more statistical interpretation of the projection.

Our first, quite simple, result concerning approximation (9) is the following.

Lemma 1. *For any supplementary statistic, A, satisfying condition* (A1), *approximation* (9) *is exact on the model subspace, which may be characterized by $\hat{D}_k = \hat{\chi}_k$ for all k.*

This result follows trivially from (10) because on the model subspace $\hat{i}_1 = \hat{j}_1$ and

$$\bar{t}' = \frac{\mathrm{d}}{\mathrm{d}\hat{\beta}}\tau(\theta(\hat{\beta})) = \hat{\Sigma}(D\hat{\theta}).$$

It turns out that the sample space derivatives needed are $\ell'(\hat{\beta}) - \ell'(\tilde{\beta})$ and $\tilde{D}_1'$, where the data point argument in all functions is $\bar{t} = \bar{t}(\hat{\beta}, A)$. One of the main points in the present paper is that these sample space derivatives can be calculated explicitly in general when approximation (9) is used, and that this approximation is sufficiently accurate.

Lemma 2. *If $\hat{P}(\hat{D}_k')$ is substituted for $\hat{D}_k'$ for all k, we obtain*

$$\ell'(\hat{\beta}) - \ell'(\tilde{\beta}) = \hat{q}^{\mathrm{T}}\hat{i}^{-1}\hat{j} \tag{13}$$

and

$$\tilde{D}_1' = \hat{S}^{\mathrm{T}}\hat{i}^{-1}\hat{j}. \tag{14}$$

The relative error due to the substitution is $O(\|\hat{\beta} - \tilde{\beta}\|\,\|A\|)$ in both cases, for any supplementary statistic with Property (A1).

Proof. To see this, consider first the log-likelihood sample space derivative in (13). Expand the log-likelihood difference in an infinite Taylor series about $\hat{\beta}$ as

$$\ell(\tilde{\beta}) - \ell(\hat{\beta}) = \hat{D}_1(\tilde{\beta} - \hat{\beta}) + \tfrac{1}{2}\hat{D}_2(\tilde{\beta} - \hat{\beta})^2 + \cdots,$$

where a suitable notation should be adopted to make this multivariate Taylor series expansion formally correct. Now differentiate the series with respect to $\hat{\beta}$ to obtain

$$\ell'(\tilde{\beta}) - \ell'(\hat{\beta}) = \hat{D}_1'(\tilde{\beta} - \hat{\beta}) + \tfrac{1}{2}\hat{D}_2'(\tilde{\beta} - \hat{\beta})^2 + \cdots,$$

because the log-likelihood derivatives with respect to the parameters vanish at the maximum values considered. Substitution of $\hat{P}(\hat{D}_k')$ from equation (12) for $\hat{D}_k'$ now leads to an infinite sum which, except for the sign reversal, may be identified with the right-hand side of expression (13). One way of doing this is to check that the expansions of the two expressions agree. Since $\hat{D}_1' = \hat{P}(\hat{D}_1')$ is an exact relation, the error from the approximation $\hat{D}_k' \approx \hat{P}(\hat{D}_k')$ in the infinite sum is $O(\|\hat{\beta} - \tilde{\beta}\|^2\|\hat{D}_k' - \hat{P}(\hat{D}_k')\|)$, which is known from Lemma 1 to be $O(n\|\hat{\beta} - \tilde{\beta}\|^2\|A\|)$, because of the smoothness of $\hat{D}_k' - \hat{P}(\hat{D}_k')$ which vanishes when $A = 0$. Since the leading term, $\hat{D}_1'(\tilde{\beta} - \hat{\beta}) = \hat{j}(\tilde{\beta} - \hat{\beta})$, is of order $n\|\hat{\beta} - \tilde{\beta}\|$, the result for the first sample space derivative follows. The result for $\bar{D}_1'$ is obtained in a similar way. □

9. Derivation of the result

In this section we sketch the proof of the result for n independent replications from a curved exponential family for which the distribution of the canonical sufficient statistic is absolutely continuous. The line of proof summarizes the development from Barndorff-Nielsen (1980; 1986; 1991), with some technical differences. The notation from the previous sections is used, in particular the n-free functions $\bar{\ell}$, i_1 and j_1 from the end of Section 6. The development assumes that the ancillary, A, has Properties (A1) and (A2) from Section 7, but several intermediate results of some general interest hold for any supplementary statistic, or assume only Property (A1). The assumptions will be explicitly stated in these cases.

9.1. APPROXIMATE CONDITIONAL DENSITY OF $\hat{\beta}$

We use f_β generically to denote the β-density of any of the statistics considered with respect to Lebesgue measure on the space in question. We start by transforming the density, $f_\beta(\bar{t})$, of $\bar{t}$ to the density of $(\hat{\beta}, A)$. It turns out to be convenient to use the variable metric $\hat{\Sigma}^{-1}$ to calculate the Jacobian of the transformation, which we may do if we multiply the density of $\bar{t}$ by $|\hat{\Sigma}|^{1/2}$, since the Riemannian measure corresponding to this metric has density $|\hat{\Sigma}|^{-1/2}$ with respect to Lebesgue measure. Determinants with respect to the variable metric are denoted $|\cdot|^*$.

The point is that we know that $\partial\bar{t}/\partial A$ belongs to the space which, in the variable metric, is orthogonal to $\hat{\Sigma}(D\hat{\theta})$, since $\hat{\beta}$ is constant when only A is changed.

Recall the definition of $\hat{P}$ from (8). The Jacobian of the transformation from $\bar{t}$ to $(\hat{\beta}, A)$ may be written

$$\left|\frac{\partial\bar{t}}{\partial(\hat{\beta}, A)}\right|^* = |\hat{P}(\bar{t}\,')|^* \left|\frac{\partial\bar{t}}{\partial A}\right|^*,$$

where the determinants on the right are generalized determinants, i.e.,

$$|M|^* = |M^{\mathrm{T}}\hat{\Sigma}^{-1}M|^{1/2} \tag{15}$$

for $M = \hat{P}(\bar{t}\,')$ or $M = \partial\bar{t}/\partial A$. Generalized determinants and related computations may be found in Tjur (1974, Section 11).

From equations (10) and (15) we see that

$$|\hat{P}(\bar{t}\,')|^* = |\,\hat{j}_1\hat{i}_1^{-1}\hat{j}_1|^{1/2} = |\,\hat{j}_1|\,|\hat{i}_1|^{-1/2}.$$

Thus the density of $(\hat{\beta}, A)$ becomes

$$f_\beta(\hat{\beta}, A) = |\hat{\Sigma}|^{1/2} f_\beta(\bar{t})|\hat{j}_1||\hat{i}_1|^{-1/2} \left| \left(\frac{\partial \bar{t}}{\partial A}\right)^{\mathrm{T}} \hat{\Sigma}^{-1} \left(\frac{\partial \bar{t}}{\partial A}\right) \right|^{1/2} \tag{16}$$

for any supplementary statistic A.

We now wish to isolate factors that mainly depend on A. The precise meaning of this is as follows.

Definition 1. *A function $h(\hat{\beta}, A)$ is called an A-function if it is constant on the model space given by $\{\bar{t} = \hat{\tau}\}$, or equivalently by $\{A = 0\}$ if A satisfies Property (A1).*

With ancillary statistics that are directed log-likelihood statistics, we can now show the following result.

Lemma 3. *For any supplementary statistic satisfying Property (A2),*

$$|\partial \bar{t}/\partial A|^{*} \text{ is an A-function.} \tag{17}$$

Proof. First note that at $A = 0$ we have

$$\hat{\Sigma}^{-1}\bar{t}' = \hat{\Sigma}^{-1}\hat{P}(\bar{t}') = (D\hat{\theta})\hat{i}_1^{-1}\hat{j}_1 = D\hat{\theta},$$

according to Lemma 1 and the fact that $\hat{i} = \hat{j}$ when $\bar{t} = \hat{\tau}$.

Since $W = 2n(\bar{\ell}(\bar{\theta}) - \bar{\ell}(\hat{\theta}))$ is assumed to be constant on level surfaces of A its derivative with respect to $\hat{\beta}$ is zero, i.e.,

$$\langle \bar{\theta} - \hat{\theta}, \bar{t}' \rangle = 0.$$

The second derivative of this equation with respect to A at 0 is the suitably symmetrized version of

$$2\left\langle \hat{\Sigma}^{-1} \frac{\partial \bar{t}}{\partial A}, \frac{\partial \bar{t}'}{\partial A} \right\rangle - \hat{\Lambda}\left(\hat{\Sigma}^{-1} \frac{\partial \bar{t}}{\partial A}, \hat{\Sigma}^{-1} \frac{\partial \bar{t}}{\partial A}, D\hat{\theta} \right) = 0, \tag{18}$$

where $\Lambda(\theta) = (\partial/\partial\theta)\Sigma(\theta)$ and we have used the notation $\hat{\Lambda}(\cdot, \cdot, \cdot)$ to denote a matrix-like multiplication of the three arguments on the three sides of the symmetric p^3-dimensional array $\hat{\Lambda}$.

This equation turns out to be what is needed to show that the derivative of $|\partial \bar{t}/\partial A|^{*}$ is zero on $A = 0$. To see this we simply calculate the derivative

$$\frac{\partial}{\partial \hat{\beta}} \left\{ \left(\frac{\partial \bar{t}}{\partial A}\right)^{\mathrm{T}} \hat{\Sigma}^{-1} \left(\frac{\partial \bar{t}}{\partial A}\right) \right\} = 2\left(\frac{\partial \bar{t}'}{\partial A}\right)^{\mathrm{T}} \hat{\Sigma}^{-1} \left(\frac{\partial \bar{t}}{\partial A}\right) - \hat{\Lambda}\left(\hat{\Sigma}^{-1} \frac{\partial \bar{t}}{\partial A}, \hat{\Sigma}^{-1} \frac{\partial \bar{t}}{\partial A}, D\hat{\theta} \right),$$

which vanishes according to (18). This proves (17). □

The final step in the rewriting of the density $f_\beta(\hat{\beta}, A)$ from (16) uses the standard saddlepoint approximation

$$f_{\bar{\theta}}(\bar{t}) = c_{k,n}|\Sigma(\bar{\theta})|^{-1/2}(1 + O(n^{-1})),$$

where $c_{k,n} = \{n/(2\pi)\}^{k/2}$, and the expansion holds uniformly for $\bar{t}$ in some bounded set. Some simple manipulations then give

$$f_\beta(\hat{\beta}, A) = c_{k,n} b(\bar{t}) |\hat{j}_1|^{1/2} \mathrm{e}^{n\{\bar{\ell}(\beta) - \bar{\ell}(\hat{\beta})\}} (1 + O(n^{-1})), \tag{19}$$

and

$$b(\bar{t}) = (|\hat{\Sigma}|^{1/2} |\Sigma(\bar{\theta})|^{-1/2})(|\hat{j}_1|^{1/2} |\hat{i}_1|^{-1/2}) \left| \frac{\partial \bar{t}}{\partial A} \right|^* \mathrm{e}^{n\{\bar{\ell}(\hat{\beta}) - \bar{\ell}(\bar{\theta})\}},$$

where we notationally allow ℓ to be a function of θ as well as of β. Note that the two factors in parentheses are A-functions, since $\bar{\theta} = \hat{\theta}$ when $\bar{t} = \hat{\tau}$. Also $|\partial \bar{t}/\partial A|^*$ has been shown to be an A-function when Property (A2) holds, and the exponential depends on $(\hat{\beta}, A)$ only through A. Thus, $b(\bar{t})$ is a function of A multiplied by an A-function. It is noteworthy that n only appears in the factor which is exactly independent of $\hat{\beta}$ – a direct consequence of Property (A2) of the supplementary statistic, A.

The marginal density of A is obtained by a Laplace-type integration over $\hat{\beta}$ using the fact that the exponent in (19) is maximal at $\hat{\beta} = \beta$. This involves some quantities defined at the point $\bar{t}_0 = \bar{t}(\beta, A)$. For fixed A, we have

$$\frac{\partial^2}{\partial \hat{\beta}^2} (\bar{\ell}(\hat{\beta}) - \bar{\ell}(\beta)) = \langle \bar{t}_0', D\theta(\beta) \rangle = n^{-1} D_1'(\beta; \bar{t}_0) = j_1(\beta; \bar{t}_0)$$

at $\hat{\beta} = \beta$. Thus, a Laplace approximation to the integral of the density in equation (19) over $\hat{\beta}$ gives

$$f_\beta(A) = c_{k-p,n} b(\bar{t}_0)(1 + O(n^{-1}))$$

uniformly in A in some bounded neighbourhood of zero. An inspection of this result shows that A, satisfying Property (A2), is a first-order ancillary statistic in the sense of having a limiting standardized distribution which is independent of β.

Division of $f_\beta(\hat{\beta}, A)$ by $f_\beta(A)$ now gives the conditional density approximation

$$\begin{aligned} f_\beta(\hat{\beta} | A) &= c_{p,n} |\hat{j}_1|^{1/2} \mathrm{e}^{n\{\bar{\ell}(\beta) - \bar{\ell}(\hat{\beta})\}} \{b(\bar{t})/b(\bar{t}_0)\}(1 + O(n^{-1})) \\ &= c_{p,n} |\hat{j}_1|^{1/2} \mathrm{e}^{n\{\bar{\ell}(\beta) - \bar{\ell}(\hat{\beta})\}} (1 + O(n^{-1}) + O(\|\hat{\beta} - \beta\| \, \|A\|)), \end{aligned} \tag{20}$$

where the omission of the factor $b(\bar{t})/b(\bar{t}_0)$ induces the relative error of order $O(\|\hat{\beta} - \beta\| \, \|A\|)$ because the factor is 1 if either $\hat{\beta} = \beta$ or $A = 0$. This is why the A-functions are collected in the factor $b(\bar{t})$. Notice that the factor omitted does not depend on n, because the ancillary has been chosen such that the exponentials in $b(\bar{t})$ and $b(\bar{t}_0)$ cancel exactly.

Formula (20) is identical to the simple version of the p^* formula from Barndorff-Nielsen (1980; 1983), using the general approximative constant $c_{p,n}$ instead of renormalizing the density. Also, several of the arguments used above, especially the omission of some factors in $b(\bar{t})$, may be recognized from the original proof in the 1980 paper.

9.2. TRANSFORMATIONS AND JACOBIANS

Having derived the density of $\hat\beta$ given A, we next transform it in two steps to the conditional density of $(\tilde\alpha, R)$. In Section 9.3 we get rid of $\tilde\alpha$, essentially by means of marginalization.

Consider the equation for the maximum likelihood estimate under the hypothesis,

$$D_1(\bar\beta) - \hat\lambda e_p = 0, \tag{21}$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier, and e_p is the vector $(0, \dots, 0, 1)^{\mathsf{T}}$. The value of λ at the maximum is denoted $\hat\lambda$.

The following lemma gives the Jacobian for the transformation from $\hat\beta$ to $(\tilde\alpha, \hat\lambda)$. It would be more correct to include A as a third component in the transformation, but since it is kept fixed throughout we omit it from the notation.

Lemma 4. *For any supplementary statistic, A, the Jacobian of the transformation from $\hat\beta$ to $(\tilde\alpha, \hat\lambda)$ is*

$$\left|\frac{\partial\hat\beta}{\partial(\tilde\alpha, \hat\lambda)}\right| = |\tilde D_1'|^{-1}|\bar j_{\alpha\alpha}|, \tag{22}$$

and the partial derivatives are

$$\frac{\partial\hat\beta}{\partial\tilde\alpha} = (\tilde D_1')^{-1}(\bar j)_{\cdot\alpha}, \tag{23}$$

where $(\bar j)_{\cdot\alpha}$ denotes the first $p-1$ columns of $\bar j$, and

$$\frac{\partial\hat\beta}{\partial\hat\lambda} = (\tilde D_1')^{-1}e_p. \tag{24}$$

Proof. Differentiation of equation (21) with respect to $\tilde\alpha$ for fixed $\hat\lambda$ gives

$$\tilde D_1'\left(\frac{\partial\hat\beta}{\partial\tilde\alpha}\right) + \tilde D_2 D\beta(\tilde\alpha) = 0,$$

where $\beta(\alpha)$ is the mapping $\alpha \mapsto (\alpha, \psi_0)$. This gives equation (23) since postmultiplication by the matrix $D\beta(\tilde\alpha)$ has the effect of picking out the first $p-1$ columns of the previous matrix.

Differentiation of equation (21) with respect to $\hat\lambda$ for fixed $\tilde\alpha$ gives

$$\tilde D_1'\left(\frac{\partial\hat\beta}{\partial\hat\lambda}\right) - e_p = 0,$$

from which equation (24) follows. The determinant is now easily computed. □

Up to this point all the approximations and derivations can be made for multivariate hypotheses with only trivial modifications. The statistic R is, however, one-dimensional by construction and the analogue of the following transformation for multivariate hypotheses is not obvious.

We wish to transform $(\tilde{\alpha}, \hat{\lambda})$ to $(\tilde{\alpha}, R)$. Since $\tilde{\alpha}$ is unchanged we only need to work out $\partial\hat{\lambda}/\partial R$.

Lemma 5. *For any supplementary statistic, A, the relevant partial derivative of the transformation from $(\tilde{\alpha}, \hat{\lambda})$ to $(\tilde{\alpha}, R)$ is*

$$\frac{\partial\hat{\lambda}}{\partial R} = R/[(\ell'(\hat{\beta}) - \ell'(\tilde{\beta}))(\tilde{D}_1')^{-1}]_p, \tag{25}$$

where $[\cdot]_p$ denotes the last coordinate of the vector. At $\hat{\beta} = \tilde{\beta}$ the limiting value replaces the expression.

Proof. The equation defining R, apart from the sign, is

$$\tfrac{1}{2}R^2 = \ell(\hat{\beta}) - \ell(\tilde{\beta}),$$

which may be differentiated with respect to R for fixed $\tilde{\alpha}$ to give

$$R = (\ell'(\hat{\beta}) - \ell'(\tilde{\beta}))\left(\frac{\partial\hat{\beta}}{\partial\hat{\lambda}}\right)\left(\frac{\partial\hat{\lambda}}{\partial R}\right).$$

Substitution of the partial derivative from (24) now immediately gives the result. □

A combination of the results above shows that for any supplementary statistic, A, we have the relation

$$\begin{aligned} f_\beta(\tilde{\alpha}, R|A) &= f_\beta(\hat{\beta}|A)\left|\frac{\partial\hat{\beta}}{\partial(\tilde{\alpha}, \hat{\lambda})}\right|\left|\frac{\partial\hat{\lambda}}{\partial R}\right| \\ &= f_\beta(\hat{\beta}|A)|\tilde{D}_1'|^{-1}|\bar{j}_{\alpha\alpha}|R/[(\ell'(\hat{\beta}) - \ell'(\tilde{\beta}))(\tilde{D}_1')^{-1}]_p, \end{aligned} \tag{26}$$

which is an exact transformation result on the domain where the transformation from $(\hat{\beta}, A)$ to $(\tilde{\alpha}, R, A)$ is one-to-one. This may be useful in, for example, transformation models where the p^* formula is known to apply, but the ancillary is different from here. For the special type of ancillaries used here the expressions for the sample space derivatives from Lemma 2 may be inserted to give a complete approximation to the conditional density of $(\tilde{\alpha}, R)$.

9.3. ELIMINATION OF NUISANCE PARAMETERS

As a final step, we need to eliminate the nuisance parameter α and its estimate $\tilde{\alpha}$. The two obvious ways of doing this are to condition on $\tilde{\alpha}$ and to marginalize from $(\tilde{\alpha}, R)$ to R. Both lead to the same result, to the order considered here, but they also both lead to the same technical difficulty which has to do with the fact that the parameter α does not disappear from the marginal or conditional density approximation for R.

Let us consider the marginalization and return to the modifications needed to resolve the difficulty mentioned above. Starting from the density (26) and the approximation (20) we approximate the integral over $\tilde{\alpha}$ by a Laplace approximation. The exponent

$$n\{\bar{\ell}(\beta) - \bar{\ell}(\hat{\beta})\} = n\{\bar{\ell}(\beta) - \bar{\ell}(\tilde{\beta})\} - \tfrac{1}{2}R^2,$$

considered as a function of $\tilde{\alpha}$ for fixed R, is maximal at $\tilde{\beta} = \beta$, or equivalently at $\tilde{\alpha} = \alpha$. Differentiation with respect to $\tilde{\alpha}$ yields

$$\frac{\partial}{\partial\tilde{\alpha}}\{\ell(\beta) - \ell(\tilde{\beta})\} = \{\ell'(\beta) - \ell'(\tilde{\beta})\}\left(\frac{\partial\hat{\beta}}{\partial\tilde{\alpha}}\right),$$

and by use of equation (23) we see that minus the second derivative at the point $\tilde{\alpha} = \alpha$ is

$$(D\beta(\tilde{\alpha}))^{\mathrm{T}}\tilde{D}_1'\left(\frac{\partial\hat{\beta}}{\partial\tilde{\alpha}}\right) = (j(\beta;\bar{t}_\alpha))_{\alpha\alpha},$$

where $\bar{t}_\alpha$ denotes the data point given by $\tilde{\alpha} = \alpha$, R and A. Also this result holds for any supplementary statistic, A.

Using the approximation (20) and the notation $\bar{R} = R/\sqrt{n}$, which is a function of $\bar{t}$ only, the Laplace integration of (26) now gives

$$f_\beta(\bar{R}|A) = \sqrt{\frac{n}{2\pi}}\mathrm{e}^{-(n/2)\bar{R}^2}g_{\hat{\alpha}}(\bar{R}|A)(1 + O(n^{-1}) + O(\|\hat{\beta} - \tilde{\beta}\|\,\|A\|)), \tag{27}$$

where

$$\begin{aligned} g_{\tilde{\alpha}}(\bar{R}|A) &= |\hat{j}|^{1/2}|\tilde{D}_1'|^{-1}|\tilde{j}_{\alpha\alpha}|^{1/2}R/[(\ell'(\hat{\beta}) - \ell'(\tilde{\beta}))(\tilde{D}_1')^{-1}]_p \\ &= \{|\hat{j}|^{-1/2}|\tilde{j}_{\alpha\alpha}|^{1/2}|\hat{i}|\,|\hat{S}|^{-1}R/[\hat{S}^{-1}\hat{q}]_p\}(1 + O(\|\hat{\beta} - \tilde{\beta}\|\,\|A\|)), \end{aligned} \tag{28}$$

and with the modification that everything should be evaluated at the data point $\bar{t}_\alpha$. This is exactly the technical difficulty referred to above, for two reasons: first, the nuisance parameter α enters the approximation; and second, the data point $\bar{t}_\alpha$ is awkward since its determination requires specification of the ancillary.

One method of overcoming this difficulty is used by Barndorff-Nielsen (1991) and by Jensen (1992). First, we simply substitute $\tilde{\alpha}$ for α in the approximation for the density. Then we continue to derive R^*, or $\tilde{R}$, from this approximation. Having derived the expression for $\tilde{R}$, we then go back to the density of $(\tilde{\alpha}, R)$ and transform this to a density of $(\tilde{\alpha}, \tilde{R})$ and again integrate out $\tilde{\alpha}$ by Laplace's method. The resulting distribution is then shown to agree with a standard normal distribution to the order considered.

The problem with the substitution of $\tilde{\alpha}$ for α in the density approximation above is that the right-hand side then depends on $\bar{t}$ through $\tilde{\alpha}$ as well as through R and A, which is unfortunate since it is supposed to approximate the density of R given A. Thus, the approximation becomes random and difficult to formalize, which is why it is more convenient to revert to the transformation to $(\tilde{\alpha}, \tilde{R})$. We shall not go through these computations which have nothing new to say, but instead continue the derivation formally from the point where formula (27) has been proved with the data point $\bar{t}$ appearing instead of $\bar{t}_\alpha$ in this formula as well as in expression (28).

Note that $g_{\tilde{\alpha}}(0|A) = 1$ and that the n's cancel in $g_{\hat{\alpha}}(\bar{R}|A)$ so that n only appears in (27) where it is explicitly written. For such density approximations, tail probabilities may be approximated either by the Lugannani–Rice method, which gives

$$\mathrm{pr}_\beta\{\bar{R} \geq \bar{r}\,|A\} \approx 1 - \Phi(\sqrt{n}\bar{r}) + \frac{\phi(\sqrt{n}\bar{r})}{\sqrt{n}\bar{r}}\,\{g_{\hat{\alpha}}(\bar{r}\,|A) - 1\},$$

or by the R^* method, giving

$$\mathrm{pr}_\beta\{\hat{R} \geq \bar{r}\,|A\} \approx 1 - \Phi\left(\sqrt{n}\bar{r} - \frac{1}{\sqrt{n}\bar{r}}\log g_{\hat{\alpha}}(\bar{r}\,|A)\right)$$

with a relative error of order $O(n^{-1} + \|\hat{\beta} - \tilde{\beta}\|\,\|A\|)$ in both cases. As mentioned, the unpleasant fact that $\tilde{\alpha}$ appears on the right in these expressions is avoided by a rigorous reformulation as in Barndorff-Nielsen (1991) or Jensen (1992).

Formula (2) for $\tilde{R}$, or the equivalent Lugannani–Rice type approximation (3), with expression (1) for $\tilde{U}$, now follows by substitution of expression (28) for $g_{\hat{\alpha}}(\bar{R}|A)$. The relative error is as stated in Section 4, because $\|\hat{\beta} - \tilde{\beta}\| = O(\|\hat{\psi} - \psi_0\|)$.

10. Discussion

The positive side of the result of the present paper is that an explicit expression has been obtained which may be of sufficient accuracy for general use, and which may therefore serve as a replacement for the usual normal-based approximations in a number of situations. There are, however, several questions, limitations and open problems, some of which are discussed below.

First of all, the present development and result deal exclusively with one-dimensional hypotheses. Generalizations to multivariate hypotheses are definitely within reach, although it is not obvious which is the best way to proceed. One way to obtain large-deviation properties for multivariate hypotheses is to use a directional approach (see Fraser and Massam 1985; Skovgaard 1988). This approach would be based on a conditioning on the direction of the score statistic from the estimate under the hypothesis, thereby effectively reducing the problem to one dimension. Other statistics than the score statistic might be considered, however, but the convenient transformation to $\hat{\lambda}$ in Section 9, which is essentially the score statistic, makes this an obvious choice. In contrast, the maximum likelihood estimate, $\hat{\psi}$, of the parameter of interest would not lead to a parametrization-invariant result.

The elimination of nuisance parameters is included here, but it is far from obvious whether the approach is sufficiently effective to deal with a large number of nuisance parameters. Presumably this will not always be the case. There are also some technical problems in connection with the elimination of these parameters, as pointed out in Section 9, and it would be nice to have a more convincing technique for this. There do not seem to be important differences between results obtained from marginalization and conditioning on the estimates, but conditioning seems more appealing because it leads to a complete elimination when the nuisance parameters are canonical parameters in an exponential family. A study such as Pierce and Peters (1992), investigating the effects of the various correction factors, might throw some light on these problems.

As discussed in relation to the example in Section 5, the method of maximum likelihood is not always reasonable, and, in fact, breaks down in the example. For variance component models most statisticians would prefer restricted maximum likelihood, which uses a marginal likelihood to estimate the variance parameters. This method does not suggest a reasonable general way of testing hypotheses about the means, however. For transformation models in general, partition of the likelihood function into marginal and conditional parts seems intuitively correct, and an adaptation of modified likelihood methods to such structured inferences would be of great practical value.

A more technical question has to do with the properties of the approximation based on $\tilde{R}$. The asymptotic behaviour stated here has been proved, although the proof was not given in all its details, but it has not been established whether the properties might be even better. In Barndorff-Nielsen and Wood (1995) it is shown that the difference between the R^*-statistic obtained from conditioning on the modified and on the unmodified directed log-likelihood is negligible to the order considered. This suggests that $\tilde{R}$ is equally valid as an approximation to the R^* obtained from the modified ancillary, as from the unmodified as used here. However, conditional statements based on the two ancillaries are not the same. The consequences of this result in the present connection are not quite clear, however.

Acknowledgement

The author wishes to thank three referees for several helpful comments on the subject material as well as its presentation. This paper was presented at the 'Conference on likelihood, asymptotic and neo-Fisherian inference' in Brixen, June 1995, at which many lectures and discussions were related to the present research topic. It is a pleasure to thank the organizers for promoting the research in this way. Discussions with Professor Don Pierce have been particularly stimulating and have affected the final version of this paper.

References

Barndorff-Nielsen, O.E. (1980) Conditionality resolutions. *Biometrika*, **67**, 293–310.

Barndorff-Nielsen, O.E. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–356.

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 Chamberlin, S.R. (1991) An ancillary invariant modification of the signed log likelihood ratio. *Scand. J. Statist.*, **18**, 341–352.

Barndorff-Nielsen, O.E. and Chamberlin, S.R. (1994) Stable and invariant adjusted directed likelihoods. *Biometrika*, **81**, 485–499.

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

Barndorff-Nielsen, O.E. and Wood, A.T.A. (1995) On large deviations and choice of ancillary for p^* and the modified directed likelihood. Research Report No. 299, Department of Theoretical Statistics, University of Aarhus.

Bleistein, N. (1966) Uniform asymptotic expansions of integrals with stationary point near algebraic singularity. *Comm. Pure Appl. Math.*, **19**, 353–370.

DiCiccio, T.J. and Martin, M.A. (1993) Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. Ser. B*, **55**, 305–316.

Fraser, D.A.S. and Massam, H. (1985) Conical tests: observed levels of significance and confidence regions. *Statist. Hefte*, **26**, 1–17.

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

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

Lugannani, R. and Rice, S.O. (1980) Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.*, **12**, 475–490.

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

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes. The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Satterthwaite, F.E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bull.*, **2**, 110–114.

Skovgaard, I.M. (1988) Saddlepoint expansions for directional test probabilities. *J. Roy. Statist. Soc. Ser. B*, **50**, 269–280.

Skovgaard, I.M. (1990) *Analytic Statistical Models*. Lecture Notes, Monograph Ser. 15. Hayward, CA: Institute of Mathematical Statistics.

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

Tjur, T. (1974) Conditional probability distributions. Lecture Notes 2. Institute of Mathematical Statistics, University of Copenhagen.

Received July 1995 and revised January 1996