The Annals of Statistics
1985, Vol. 13, No. 2, 534–551

A SECOND-ORDER INVESTIGATION OF ASYMPTOTIC ANCILLARITY

BY IB M. SKOVGAARD

Royal Veterinary and Agricultural University, Copenhagen

The paper deals with approximate ancillarity as discussed by Efron and Hinkley (1978). In the multivariate i.i.d. case we derive the second-order Edgeworth expansion of the MLE given a normalized version of the second derivative of the log-likelihood at its maximum. The expansion agrees with the one derived by Amari (1982a) for curved exponential families, but holds for any family satisfying the regularity conditions given in the paper. It is shown that the Fisher information lost by reducing the data to the MLE is recovered by the conditioning, and it is sketched how the loss of information relates to the deficiency as defined by LeCam. Finally, we investigate some properties of three test statistics, proving a conjecture by Efron and Hinkley (1978) concerning the conditional null-distribution of the likelihood ratio test statistic, and establishing a kind of superiority of the observed Fisher information over the expected one as estimate of the inverse variance of the MLE.

1. Introduction. The purpose of this paper is to investigate some properties related to the conditioning on asymptotic ancillaries as proposed by Efron and Hinkley (1978). Since exact properties are hard to derive in general, the investigation is carried out in terms of second-order asymptotic distributions, i.e., including the $n^{-1/2}$ terms in the asymptotic expansions. It turns out that *first-order asymptotics fail to discriminate between the conditional approach and the usual (marginal) approach.* Emphasis will be on the results, since the techniques used to prove these are largely well-known, but in Section 7 we shall sketch the ideas of the proofs.

Since the arguments for conditioning on (approximately) ancillary statistics are outlined in Efron and Hinkley (1978), we shall not discuss the issue at length, but merely give an example, essentially based on Pierce (1975), illustrating the advantages of this approach.

EXAMPLE 1.1 Let $(\bar{X}, \bar{Y})$ be the average of n independent two-dimensional normal variables, each with the identity matrix as covariance and with mean $\mu(\beta) \in \mathbb{R}^2$, where β is a real parameter, and μ is some smooth function. For each β, let L_β denote the line through $\mu(\beta)$ orthogonal to the tangent at $\mu(\beta)$. If $\hat{\beta}$ is the maximum likelihood estimator of β, then the observation $(\bar{x}, \bar{y})$ must be on the line $L_{\hat{\beta}}$; see Efron (1978) for further geometrical details. If n is large, we may for inferential purposes approximate $\mu(\beta)$ locally by a segment of a circle (see Figure 1). Let P denote the center of this circle; then the lines L_β will for β near to $\hat{\beta}$ approximately go through P. Now, if we want a confidence interval for β, a

Received June 1981; revised September 1984.

AMS 1980 *subject classifications.* Primary 62E20, secondary 62F12.

Key words and phrases. Ancillarity, deficiency, Edgeworth expansions, loss of information, maximum likelihood, observed Fisher information, second-order asymptotics, Wald's test.

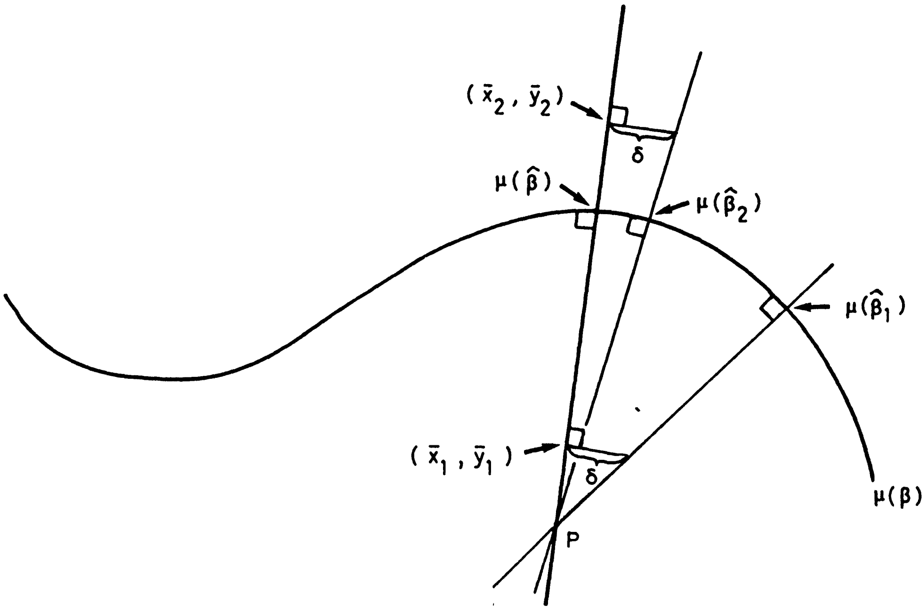

FIG. 1. *Accuracy of the maximum likelihood estimate. The sensitivity of the estimate due to a displacement δ of the observation depends on the distance of $(\bar{x}, \bar{y})$ to P.*

common method will be to "center" this interval at $\hat{\beta}$ and let the length be approximately proportional to the standard deviation of $\hat{\beta}$, disregarding the position of $(\bar{x}, \bar{y})$ on the line $L_{\hat{\beta}}$. However, if the observation is $(\bar{x}_1, \bar{y}_1)$, a displacement of this by an amount δ orthogonal to $L_{\hat{\beta}}$ would change the estimate from $\hat{\beta}$ to $\hat{\beta}_1$, whereas, if the observation is $(\bar{x}_2, \bar{y}_2)$, a similar displacement would only change the estimate to $\hat{\beta}_2$. This suggests that the "accuracy" of the estimate is somehow increasing with the distance of $(\bar{x}, \bar{y})$ from the center P. It may be noted that confidence intervals constructed using the likelihood ratio test would certainly reflect this fact. This may be seen by noticing that if the observation is near the center P, the distance to the curve $\mu(\beta)$ is almost constant in the neighbourhood of $\mu(\hat{\beta})$. In more general examples, similar considerations hold, but the geometrical picture is not equally obvious. □

The example shows that the estimator may not be sufficiently informative. A natural way to try to improve it is to look for an ancillary statistic and replace the marginal distribution of the estimator by its conditional distribution given the ancillary statistic.

Several suggestions of ancillaries capturing some of this additional information have been put forward (see e.g. Barndorff-Nielsen, 1980), but except for one they are related to exponential models or other specific classes of models, e.g., translation models. The remaining one is essentially the second derivative of the log-likelihood function at its maximum. This idea goes back to Fisher, but was suggested by Efron and Hinkley (1978) in more explicit form. A problem is that this second derivative may contain substantial information; however, after a suitable normalization, it will be asymptotically ancillary.

Since this conditional distribution can rarely be calculated exactly, we shall derive an asymptotic expansion of its distribution, which may serve as an analogue of the more familiar expansions of the marginal distribution of the MLE. To distinguish between the two methods, one has to compute the second-order asymptotic expansions, i.e., include the $n^{-1/2}$ terms of the distributions in the case of n replications.

The resulting expansion for the conditional distribution is given in Section 4. The expansion is of the Edgeworth type determined by the first three asymptotic central moments of the conditional distribution, and it is seen that *only the variance is changed compared to the unconditional expansion.*

This expansion may formally be derived from Expansion (6.5) in Amari (1982a), which is more general in the sense that it is not restricted to maximum likelihood estimators. The derivation is, however, restricted to curved exponential families, although the result is not, in the sense that the quantities in the expansion are defined outside this framework. The ancillary statistic in Amari's paper is any statistic which together with the estimator is minimally sufficient, and it may therefore be of higher dimension than the Efron-Hinkley ancillary. A heuristic argument suggesting the equivalence of the two approaches is as follows: The maximum likelihood estimator together with the second derivative of the log-likelihood function at its maximum is a sufficient statistic to second order (see Section 5). Therefore the model can be approximated by a curved exponential family generated by these two statistics, and within this family, the ancillaries of Amari's paper will coincide (locally) with the standardized versions of the second derivatives of the log-likelihood, i.e., the Efron-Hinkley ancillaries.

Knowledge of this expansion makes the conditional approach feasible, but provides no justification for the method. An investigation in this direction is given in Section 5 in terms of (loss of) *Fisher information.* It is shown that the average Fisher information contained in the conditional distribution differs from that of the whole set of data only by an amount which tends to zero as the number of replications increases, whereas the corresponding deficit for the unconditional distribution converges to a fixed quantity depending on the curvature of the model (see Efron, 1975).

A more operationally meaningful way of defining the loss of information involved in a data reduction is in terms of the *deficiency* (LeCam, 1964), which roughly speaking measures how well any test based on the full data set can be mimicked using only the reduced data. The deficiencies for the reductions of the data to the MLE and to the MLE supplemented by the ancillary statistic are also investigated in Section 5, where it is shown as a general result that *the deficiency is bounded by an amount proportional to the square root of the relative loss of Fisher information.* Thus the results concerning loss of Fisher information may be interpreted as providing bounds for the deficiencies for the various reductions.

The variance of the second-order conditional distribution as calculated in Section 4 is, when evaluated at the MLE, equal to the observed Fisher information. Hence, a natural consequence would be to replace the (expected) Fisher information in the Wald test statistic by the observed Fisher information (see Efron and Hinkley, 1978). A comparison of the behaviors of these two test statistics compared to the likelihood ratio test statistic is carried out in Section 6 in two ways. First it is shown that the null-distribution given the asymptotic ancillary statistic converges more rapidly (namely as $O(n^{-1})$) towards its chi-squared limit if the observed Fisher information is used than if the expected one is used, when the convergence rate is $O(n^{-1/2})$. This was conjectured by Efron and Hinkley (1978), and has been shown by Peers (1978) for the one-parameter case. Next, it is shown that the use of the observed information is also superior in the sense that it provides a better stochastic approximation to the likelihood ratio test than the usual Wald test, when null-distributions are considered.

Proofs have been carried through for the case of n replications under the assumptions given in Section 7, which also contains sketches of the proofs, but we shall not give these in detail since they are based mainly on well-known techniques. Some basic ideas of the proofs are, however, given along with the results.

To introduce the results, some examples are given in Section 3; in particular, Example 3.2 should give an impression of the kind and amount of computations required for applications.

2. Preliminaries.

Let $X_1, \cdots, X_n$ be independent identically distributed random variables on some measurable space, and suppose that the distribution of X_i is a member of a family $\{P_\beta, \beta \in B \subseteq \mathbb{R}^p\}$, where B is some subset of $\mathbb{R}^p$. We assume that the family is dominated by some measure μ on $\mathbb{R}^p$, and let $f(x, \beta)$ denote the densities. We also assume that the conditions of Section 7 are fulfilled, these essentially being various kinds of smoothness conditions.

For each $\beta \in \operatorname{int}(B)$ define

$$D_j(\beta) = (1/n) \textstyle\sum_{i=1}^n D^j \log f(X_i, \beta), \quad E_j(\beta) = E_\beta\{D_j(\beta)\}$$

where $D^j \log f(X_i, \beta)$ denotes the j-sided array of jth derivatives with respect to β. Also, we define the joint cumulants of these derivatives by

$$\chi_{j\cdots k}(\beta) = \operatorname{cum}_\beta\{D^j \log f(X_i, \beta), \cdots, D^k \log f(X_i, \beta)\}$$

which is of dimension $p^j \cdots p^k = p^{j+\cdots+k}$. In particular the (expected) *Fisher information* per observation is

$$I(\beta) = -E_2(\beta) = \chi_{11}(\beta),$$

whereas the *observed information* per observation is

$$\hat{J} = J(\hat{\beta}) = -D_2(\hat{\beta}).$$

For convenience we shall use the following notational conventions. If the argument β is omitted, a fixed point β_0 (the "true" parameter) is understood, while a circumflex indicates evaluation at the maximum likelihood estimate $\hat{\beta}$, e.g. $I = I(\beta_0)$ and $\hat{I} = I(\hat{\beta})$. Dependence on n is usually not explicitly indicated.

For multiplying vectors, matrices and arrays we shall sometimes use coordinates to clarify definitions and results; otherwise we use some notation that is most easily explained by a few examples given below. When $v = (v_k)$, $k = 1, \cdots, r$ is a vector, $M = (m_{jk})$ is a $q \times r$ matrix, and $A = (a_{ijk})$ is a $p \times q \times r$ array, then:

$$\begin{array}{lll} Mv = \sum_k m_{jk} v_k & & [\text{vector}] \\ M(v^2) = \sum_{j,k} m_{jk} v_j v_k, & \text{if} \quad q = p & [\text{scalar}] \\ Av = (\sum_k a_{ijk} v_k) & & [\text{matrix}] \\ A(v^3) = \sum_{ijk} a_{ijk} v_i v_j v_k & & [\text{scalar}]. \end{array}$$

Finally we use $\langle\ ,\ \rangle$ to denote the inner product, e.g.

$$\langle v, w \rangle = \textstyle\sum_k v_k w_k, \quad \langle M, N \rangle = \sum_{j,k} m_{jk} n_{jk},$$

where w is an r-vector and N is a $q \times r$-matrix (as is M). The coordinates are collected in an obvious manner, e.g.

$$(\chi_{22})_{ij,kl} = \operatorname{Cov}((d^2/d\beta_i d\beta_j)\log f(X_1, \beta), (d^2/d\beta_k d\beta_l)\log f(X_1, \beta)).$$

The *Efron-Hinkley ancillary statistic* A is defined as a standardized version of the observed information, i.e.

$$A = \sqrt{n}\, \hat{F}^{-1/2}(\hat{J} - \hat{I}) \in \mathbb{R}^d, \tag{2.1}$$

where $F(\beta)$, defined as

$$F(\beta)_{ij,kl} = (\chi_{22})_{ij,kl} - \textstyle\sum_{\alpha,\gamma} (\chi_{21})_{ij,\alpha} (I^{-1})_{\alpha\gamma} (\chi_{12})_{\gamma,kl} \tag{2.2}$$

is the asymptotic variance of $\sqrt{n}\ (\hat{J} - \hat{I})$. Here d is the rank of F, and $F^{-1/2}$ is defined as any smooth "square-root" of F, i.e.

$$(F^{-1/2})' F^{-1/2} = 1_d (\text{the identity matrix on } \mathbb{R}^d),$$

which is the asymptotic variance of A. It may be noted that F is the residual variance of D_2 after regression on D_1.

In the one-dimensional case $F = (\gamma I)^2$, where γ is the statistical curvature defined in Efron (1975). In the multidimensional case F is the square of the second fundamental form in a differential geometry corresponding to that used by Efron (1975); see Reeds (1975) and Madsen (1979, page 24).

3. Examples.

EXAMPLE 3.1. *Non-linear normal regression.* Let $X_1, \cdots, X_n$ be independent normal random variables with variance σ^2 and expectation vector $\mu(\beta)$, where $\beta \in \mathbb{R}^p$ and $\mu: \mathbb{R}^p \to \mathbb{R}^n$ is a known function. The variance σ^2 is considered known, since this is notationally more convenient and has no influence on the estimate of β and its distribution. These models do not fit into the i.i.d. framework, but the asymptotics $\sigma^2 \to 0$ corresponds to independent replications of the entire experiment (since in that case $\sigma^2/n \to 0$ as $n \to \infty$), and is furthermore reasonable in cases where σ is small.

The information matrix is

$$I = \sum_{i=1}^{n} (D\mu_i)'(D\mu_i)/\sigma^2, \quad \mu_i(\beta) = E_\beta X_i \tag{3.1}$$

and the difference $\hat{J} - \hat{I}$ between the observed and expected Fisher information at $\hat{\beta}$ is

$$\Delta = \hat{J} - \hat{I} = -\sum_{i=1}^{n} (X_i - \hat{\mu}_i)D^2\hat{\mu}_i/\sigma^2, \tag{3.2}$$

where $D^2\hat{\mu}_i$ is the matrix of second derivatives of μ_i at $\hat{\beta}$. The ancillary statistic A is a normalized version of the statistic Δ, but it need not be calculated in applications.

The asymptotic unconditional distribution of $\sqrt{n}(\hat{\beta} - \beta)$ is the well-known normal approximation with variance matrix I^{-1}, and the second-order expansion of this distribution is given by an Edgeworth expansion with the same variance, but with a bias and third cumulant of order $O(n^{-1/2})$, which are easily calculated, see, e.g., Skovgaard (1980b).

The conditional second-order distribution is obtained from the unconditional one merely by replacing the inverse variance matrix I by $I + \Delta$ (see (4.12)), or equivalently we may write the variance matrix itself as

$$\operatorname{Var}(\hat{\beta} \mid A) \approx I^{-1}\{1_p + (\sum_{i=1}^{m} (x_i - \hat{\mu}_i)D^2\hat{\mu}_i)I^{-1}\}. \tag{3.3}$$

It should be noted that this depends on the true parameter value β_0 through I^{-1}, and that if the common practice of replacing I by $\hat{I}$ is adopted, the distributional approximation will in general only be of first order. However, if $\hat{J}^{-1}$ is used as an approximation to the variance (or equivalently I replaced by $\hat{I}$ in (3.3)), this will be superior to the usual approximation $\hat{I}^{-1}$ for testing an hypothesis about β, in two specific ways stated in Theorem 6.1 and Theorem 6.2.

If the random variables $X_1, \cdots, X_n$ are correlated with covariance matrix $\sigma^2\Sigma$, where Σ is known, the same results hold, except that the sums in (3.1), (3.2) and (3.3) are replaced by inner products with respect to Σ^{-1}. □

EXAMPLE 3.2. To be more specific we shall consider an example of nonlinear normal regression. Consider a *logistic growth function* of the form $e^{\alpha(t-\gamma)}/\{1 + e^{\alpha(t-\gamma)}\}$, where t is time (the independent variable), α and γ are unknown parameters and the growth is supposed to be scaled, such that the limiting "size" is 1. We shall consider this model in logarithmic scale, i.e. we assume that $X_1, \cdots, X_n$ are independent normal variables with variance σ^2, and with expectations

$$EX_i = \mu_i(\alpha, \gamma) = \alpha(t_i - \gamma) - \log\{1 + e^{\alpha(t_i-\gamma)}\}, \tag{3.4}$$

where $t_1, \cdots, t_n$ are known time points.

The information matrix is

$$I = \sum_{i=1}^{n} (D\mu_i)'D\mu_i/\sigma^2 = \sum_{i=1}^{n} M_i/(d_i^2\sigma^2) \tag{3.5}$$

where $d_i = 1 + e^{\alpha(t_i-\gamma)}$ and

$$M_i = \begin{pmatrix} (t_i - \gamma)^2 & -\alpha(t_i - \gamma) \\ -\alpha(t_i - \gamma) & \alpha^2 \end{pmatrix}$$

and the difference $\Delta = \hat{J} - \hat{I}$ equals (cf. (3.2))

$$\Delta = \textstyle\sum_{i=1}^{n} (x_i - \hat{\mu}_i)e^{\hat{\alpha}(t_i-\hat{\gamma})}\hat{M}_i/(\hat{d}_i^2\sigma^2). \tag{3.6}$$

To derive the conditional distribution of $(\hat{\alpha} - \alpha, \hat{\gamma} - \gamma)$ one needs to compute the bias and the third cumulant of the second-order approximation to this vector. The general formulae are well-known, and in this example the bias is

$$\frac{1}{2}\hat{I}^{-1} \textstyle\sum_{i=1}^{n} \begin{pmatrix} t_i - \hat{\gamma} \\ -\hat{\alpha} \end{pmatrix} e^{\hat{\alpha}(t_i-\hat{\gamma})}\text{trace}\{\hat{M}_i\hat{I}^{-1}\}/d_i^3, \tag{3.7}$$

and the third cumulant is of similar complexity, but more cumbersome to write, because it is a 2 × 2 × 2 array. The conditional distribution of $(\hat{\alpha}, \hat{\gamma})$ is now given by (4.11) in terms of these cumulants. □

EXAMPLE 3.3. This example is the one from the end of the paper by Hinkley (1980). Let (Y_i, Z_i) be i.i.d. bivariate normal variables with Z_i distributed as $N(\theta_1, 1)$ and $Y_i = \theta_2 Z_i + \varepsilon_i$, where ε_i is $N(0, 1)$. By simple computations we get

$$\hat{\theta}_1 = \bar{Z} = \textstyle\sum Z_i/n, \qquad \hat{\theta}_2 = \sum Z_i Y_i/\sum Z_i^2$$

$$I(\theta) = \text{diag}(1, 1 + \theta_1^2), \quad \hat{J} = \text{diag}(1, \textstyle\sum Z_i^2/n).$$

Since $\hat{J} - \hat{I} = \text{diag}(0, \sum (Z_i - \bar{Z})^2/n - 1)$ has one-dimensional support, we only compute the corresponding element of F, i.e. $F_{2222} = 2$, and define (see (2.1))

$$A = (\textstyle\sum (Z_i - \bar{Z})^2 - n)\sqrt{2n}.$$

A is seen to be exactly ancillary and $(\hat{\theta}, A)$ is sufficient. $\hat{\theta}_1$ is independent of A, and since the conditional distribution of $\hat{\theta}_2$ given $Z_1, \cdots, Z_n$ is $N(\theta_2, (\sqrt{2n}A + n(1 + \hat{\theta}_1^2))^{-1})$, it follows, that to second order the conditional distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ given $A = a$ is normal with mean zero and

$$V\{(\hat{\theta}_1, \hat{\theta}_2)\} \sim \text{diag}(n^{-1}, (\sqrt{2n}a + n(1 + \theta_1^2))^{-1}) = n^{-1}(\hat{J} + I - \hat{I})^{-1}$$

in agreement with (4.12). Also, if L is the likelihood ratio statistic, $W = (\hat{\theta} - \theta)'\hat{I}(\hat{\theta} - \theta)$ is the Wald test statistic, and $W_c = (\hat{\theta} - \theta)\hat{J}(\hat{\theta} - \theta)$ is the modified Wald test statistic obtained by using the observed information J instead of I, then, as noted by Hinkley, $L = W_c = n(\bar{Z} - \theta_1)^2 + (\sum Z_i\varepsilon_i)^2/\sum Z_i^2$ is exactly distributed as X_2^2, whereas W deviates from this by an amount of order $n^{-1/2}$ (cf. Theorem 6.1 and 6.2).

4. Expansion of the conditional distribution.

In this section we shall expand the conditional distribution of $Z = \sqrt{n}(\hat{\beta} - \beta_0)$ given A under the distribution P_{β_0}. It is not hard to prove that to first order Z and A are asymptotically independent. Thus to obtain any interesting results, we must carry the expansion to second order, i.e., include the $n^{-1/2}$ terms. The first step is to expand the joint distribution of (Z, A). This is done in the following three steps:

(i) the second-order (stochastic) Taylor-series expansion of (Z, A) in terms of the derivatives of the log-likelihood at β_0 is computed;

(ii) the first three joint cumulants of these approximating polynomials are computed—these will be functions of the E's and χ's;

(iii) by insertion of these cumulants into the general formula for the Edgeworth approximation, the joint distribution is obtained. Since the expansion obtained in this way is the basis of all our results, we shall state it in detail in Theorem 4.1 below.

We shall refer to the cumulants of the approximating distribution as κ_z resp. κ_a for the first cumulants (the means) of Z resp. A, κ_{zzz} the third cumulant of Z, κ_{zza} the mixed third cumulant of Z, Z, A, etc. The cumulants needed in the expansion of the conditional distribution are given by

$$\langle \kappa_z, Iz \rangle = -\tfrac{1}{2}\langle I^{-1}, \chi_{111}(z) + \chi_{21}(z)\rangle/\sqrt{n} \tag{4.1}$$

$$\kappa_{zzz}((Iz)^3) = -(2\chi_{111}(z^3) + 3\chi_{12}(z^3))/\sqrt{n} \tag{4.2}$$

$$\kappa_{zza}(Iz, Iz, a) = -\langle F(z^2), (F^{-1/2})'(a)\rangle/\sqrt{n} \tag{4.3}$$

$$\kappa_{zaa} = 0 \tag{4.4}$$

for all $z \in \mathbb{R}^p$, a $\in \mathbb{R}^d$.

THEOREM 4.1. *Under Conditions* 7.1 *we have the following local expansions for any* $c > 0$:

$$\sup\{|g_n(z, a) - \gamma_n(z, a)|;\ \|z, a\|^2 \leq c \log n\} = O(n^{-1}) \tag{4.5}$$

$$\sup\{|h_n(a) - \xi_n(a)|;\ \|a\|^2 \leq c \log n\} = O(n^{-1}), \tag{4.6}$$

where $\|\cdot\|$ *denotes the Euclidean norm*, g_n *and* h_n *are the densities of* (Z, A) *and* A, *and*

$$\begin{aligned}\gamma_n(z, a) = {} & (2\pi)^{-(p+d)/2}(\det I)^{1/2}\{\exp -\tfrac{1}{2}(I(z^2) + \langle a, a\rangle)\}\\ & \times (1 + \langle \kappa_z, Iz\rangle + \langle \kappa_a, a\rangle + \tfrac{1}{6}\kappa_{zzz}((Iz)^3) + \tfrac{1}{6}\kappa_{aaa}(a^3)\\ & + \tfrac{1}{2}\kappa_{zza}(Iz, Iz, a) - \tfrac{1}{2}\langle I, \kappa_{zzz}(Iz)\rangle - \tfrac{1}{2}\langle I, \kappa_{zza}(a)\rangle\\ & - \tfrac{1}{2}\operatorname{trace}\{\kappa_{aaa}(a)\}),\end{aligned} \tag{4.7}$$

$$\begin{aligned}\zeta_n(a) = {} & (2\pi)^{-d/2}\exp\{-\tfrac{1}{2}\langle a, a\rangle\}\\ & \cdot (1 + \langle \kappa_a, a\rangle + \tfrac{1}{6}\kappa_{aaa}(a^3) - \tfrac{1}{2}\operatorname{trace}\{\kappa_{aaa}(a)\}),\end{aligned} \tag{4.8}$$

are the Edgeworth approximations to the two densities, in which d is the dimension of A.

To clarify the meaning of the notation, we shall give formula (4.7) for the case where Z and A (and hence all the κ's) are one-dimensional. We then have

$$\begin{aligned}\gamma_n(z, a) = {} & (2\pi)^{-(p+d)/2}I^{-1/2}\{\exp -\tfrac{1}{2}(Iz^2 + a^2)\}\\ & \times (1 + \kappa_z Iz + \kappa_a a + \tfrac{1}{6}\kappa_{zzz}I^3z^3 + \tfrac{1}{6}\kappa_{aaa}a^3\\ & + \tfrac{1}{2}\kappa_{zza}I^2z^2a - \tfrac{1}{2}\kappa_{zzz}I^2z - \tfrac{1}{2}I\kappa_{zza}a - \tfrac{1}{2}\kappa_{aaa}a),\end{aligned} \tag{4.9}$$

which is, in fact, not much different from (4.7).

It is seen from (4.8) combined with the fact that the first and third cumulant (κ_a and κ_{aaa}) of A are both $O(n^{-1/2})$, that A is not, in general, second-order ancillary in the sense that the second-order distribution is independent of β_0; but it is locally second-order ancillary in the sense of Cox (1980). This means that in any set of the form $\{\beta;\ \|\beta - \beta_0\| \leq c/\sqrt{n}\}$ with $c > 0$ fixed, the distribution of A is constant except for terms of order $O(n^{-1})$. This is the property that turns out to be important to avoid loss of information (see Section 5).

The expansion of the conditional distribution of Z given A may now be obtained by dividing $\gamma_n(z, a)$ by $\zeta_n(a)$, although the proof requires further expansion than given in Theorem 4.1.

THEOREM 4.2. *Under Conditions* 7.1 *we have the following expansion of the conditional distribution of* $Z = \sqrt{n}\ (\hat{\beta} - \beta_0)$ *given* $A = a$,

$$P\{Z \in B \mid A = a\} = \int_B \eta_n(z \mid a)\ dz + O(n^{-1}) \tag{4.10}$$

uniformly over all Borel sets $B \in \mathbb{R}^p$ *and* $\|a\|^2 \leq (2 + \alpha) \log n$, *for some* $\alpha > 0$, *where*

$$\begin{aligned}\eta_n(z \mid a) = {} & (2\pi)^{-p/2}\{\det(I + F^{1/2}(a))\}^{1/2}\exp\{-½(I + F^{1/2}(a))(z^2)\} \\ & \times \{1 + \langle \kappa_z, Iz \rangle + ⅙\kappa_{zzz}((Iz)^3) - ½\langle I, \kappa_{zzz}(Iz)\rangle\}\end{aligned} \tag{4.11}$$

is the second-order expansion of the condition density.

REMARK. It is important to note that the event $\{\|A\|^2 \leq (2 + \alpha)\log n\}$ has probability $1 - O(n^{-1})$, so that Theorem 4.2 together with (4.6) implies, that

$$P\{Z \in B\} = \int_{\|a\|^2 \geq (2+\alpha)\log n} \zeta_n(a) \int_B \eta_n(z \mid a)\ dz\ da + O(n^{-1}).$$

A local expansion of the conditional density of Z given A holding uniformly only on a bounded set, would not suffice to prove this, and in this sense the result would be incomplete.

There are some points worth noting about the moments of η_n. The first and third moment are (to second order) independent of a, and the same as in the unconditional second-order expansion, whereas the variance depends on a. The theorem says nothing about the conditional moments of the exact distribution, but if these are to be used as descriptive quantities of the distribution, then rather than expanding these, it is the moments of the approximating distribution that are relevant.

To second order we have

$$\tilde{V}(Z \mid A = a)^{-1} = I + n^{-1/2}F^{1/2}(a) \sim \hat{J} + I - \hat{I}, \tag{4.12}$$

where $\tilde{V}$ is the variance of the approximate distribution. Thus it is seen that if the common practice of inserting the estimate $\hat{\beta}$ for the unknown parameter β_0 is used in Formula (4.12), one arrives at the observed Fisher information $\hat{J}$ as an estimate of the inverse variance in the approximate conditional distribution, as noted by Efron and Hinkley (1978) and Amari (1982a), Formula (6.12). If, however, this approximation is used, the distributional approximation is no longer of second order, and it is questionable whether anything has been gained compared to the usual unconditional first-order approximation.

In the special case when the derivative of $I(\beta)$ at β_0 vanishes, as is the case of the translation models considered in Efron and Hinkley (1978), then the approximation

$$\tilde{V}(Z \mid A = a) \sim \hat{J}^{-1}$$

will, however, lead to a second-order approximation to the conditional distribution. In contrast the bias and the third cumulant of Z may in general be replaced by estimates obtained through evaluation at $\hat{\beta}$ without changing the order of magnitude of the distributional approximation in (4.10).

5. Recovery of information. Fisher's main reason for considering ancillaries and, more specifically, conditional distributions given ancillaries was that by the reduction to a single statistic, such as the MLE, one might lose a certain amount of (Fisher) information, which might be "recovered" by a conditional approach.

The total amount of Fisher information in the experiment is $nI(\beta_0) = \inf(X)$, say, where $X = (X_1, \cdots, X_n)$. In general we let $\inf(T)$ denote the Fisher information (at β_0) contained in an experiment where only T is observed. Also, we shall consider the information $\inf(T \mid A = a)$ in the experiment, where $A(X) = a$ is fixed and T is observed, and its expected value is $\inf_A(T) = E\{\inf(T \mid A)\}$. The well-known identity $\inf(T) = \inf(X) - E\{\mathrm{Var}\{D \log f(X; \beta_0) \mid T\}\}$, see e.g. Fisher (1925), is useful in computing $\inf(T)$. It is well-known, see Fisher (1925), that $\inf(X) - \inf(\hat{\beta})$ tends to a finite limit as $n \to \infty$, which Efron (1975) identified as $\gamma^2 I$ in the one-dimensional case, where γ is the statistical curvature of the model at β_0. The following theorem shows that this information lost by the reduction of X to $\hat{\beta}$ is indeed recovered by conditioning by A as defined in (2.1).

THEOREM 5.1. *Under Conditions* 7.1 *we have*

$$\inf(X) - \inf(\hat{\beta}) = F(\cdot, I^{-1}, \cdot) + O(n^{-1}) \tag{5.1}$$

$$\inf(X) - \inf(\hat{\beta}, A) = O(n^{-1}) \tag{5.2}$$

$$\inf(A) = O(n^{-1}) \tag{5.3}$$

$$\inf_A(\hat{\beta}) = \inf(X) - O(n^{-1}) \tag{5.4}$$

where $F(\cdot, I^{-1}, \cdot)$ is the matrix with entry (i, j) given by $\sum_{k,l} F_{ik,lj}(I^{-1})_{kl}$

Note that (5.4) follows from (5.2) and (5.3), since $\inf_A(\hat{\beta}) = \inf(\hat{\beta}, A) - \inf(A)$. Formal proofs of (5.1) and (5.2) go back to Fisher (1925), whereas Rao (1961) gave a strict proof of (5.1) in the multinomial case; see Efron (1975), Section 9, for further discussion and references. Strict proofs may be given under weaker assumptions than those of Section 7, but we shall not elaborate on this point.

If one does not believe, as Fisher seemed to, that the (Fisher) information is an absolute measure of information, then it would be natural to look for other interpretations or implications of Theorem 5.1 and similar results; see LeCam (1975). A reasonable possibility would be to measure the information lost in the reduction from $X = (X_1, \cdots, X_n)$ to $T = T_n(X)$ by the *deficiency* of the experiment $(Q_\beta, \beta \in B)$ with respect to the experiment $(P_\beta, \beta \in B)$, where Q_β is the distribution of T, as defined by LeCam (1964). The deficiency is defined as the "distance" between the original distribution and the best "reconstruction" of it based on T by a randomization which is independent of the parameter. As the distance is used the maximal difference in probability over all sets, and the supremum over all parameters is the deficiency. This is an intuitively appealing measure, because it tells something about the probabilistic performance of the two experiments, namely how well *any* test based on X can be approximated by use of T only.

In agreement with LeCam (1956) and Michel (1978) we shall use a slight modification of the deficiency by restricting attention to compact sets of parameter values. Define

$$
\begin{aligned}
\delta_K(T, X) &= \inf_\Pi \sup_{\beta \in K} \tfrac{1}{2} \, \| P_\beta - \Pi Q_\beta \| \\
&= \inf_\Pi \sup_{\beta \in K} \sup_A \{ | P_\beta(A) - (\Pi Q_\beta)(A) | \}, \quad K \subseteq \mathbb{R}^P \text{ compact}
\end{aligned}
\tag{5.5}
$$

where Π varies over the class of Markov-kernels and A over all measurable sets. Except for minor technical differences concerning the class of kernels Π this is the deficiency of $(Q_\beta, \beta \in K)$ with respect to $(P_\beta, \beta \in K)$. Attention is restricted to compact sets $K \subseteq B$, since uniform approximation over B can hardly be obtained in general. Notice that $\delta_K(T, X) = 0$ if T is sufficient.

Let us now assume, that $\hat{\beta}$ is a function of T, although another first-order efficient estimator might do as well as $\hat{\beta}$, and let us define $\Pi = P^t_{\hat{\beta}}$, i.e. the (ΠQ_β)-conditional distribution of X given $T = t$ is $P^t_{\hat{\beta}}$, where P^t_β is the P_β-conditional distribution of X given $T = t$. We shall give a formal proof that $\delta_K(T, X)$ *is asymptotically bounded by the maximum over K of the square root of the relative loss of Fisher information.* More precisely

$$
\| P_\beta - \Pi Q_\beta \| \leq \sqrt{p}(\text{trace } R_\beta(T))^{1/2}(1 + o(1))
\tag{5.6}
$$

where p is the dimension of β and $R_\beta(T) = \inf(X)^{-1}(\inf(X) - \inf_\beta(T))$ is the relative loss of Fisher information.

Let $f^t(x; \beta)$ denote the density of P^t_β with respect to μ. The proof of (5.6) then goes as follows

$$
\begin{aligned}
\| P_\beta - \Pi Q_\beta \| &= \int \int | f^t(x; \beta) - f^t(x; \hat{\beta}) | \, d\mu(x) \, dQ_\beta(t) \\
&\sim \int \int | (D_\beta \log f^t(x; \beta))(\hat{\beta} - \beta) | \, dP_\beta(x) \, dQ_\beta(t) \\
&\leq \int \int \| I(\beta)^{-1/2}(D_\beta \log f^t(x; \beta)) \| \\
&\quad \cdot \| I(\beta)^{1/2}(\hat{\beta} - \beta) \| \, dP_\beta(x) \, dQ_\beta(t)
\end{aligned}
$$

$$\begin{aligned} &\le (E_\beta\{(nI(\beta))(\hat\beta - \beta)^2\})^{1/2} \\ &\quad \cdot (E_\beta\{\langle (nI(\beta))^{-1}, \inf_\beta(X \mid T)\rangle\})^{1/2} \\ &\le \sqrt{p}(E_\beta\{\mathrm{trace}((ni(\beta))^{-1}\inf_\beta(X \mid T))\})^{1/2} \\ &= \sqrt{p}(\mathrm{trace}\{\inf_\beta(X)^{-1}(\inf_\beta(X) - \inf_\beta(T))\})^{1/2} \end{aligned}$$

where the second inequality follows from Hölders inequality.

Using this result together with Theorem 5.1 we see, that $\delta_K(\hat\beta, X) = O(n^{-1/2})$ and $\delta_K((\hat\beta, A), X) = O(n^{-1})$, which may be viewed as a special case of the result in Michel (1978), where the statistics of the form $T = (\hat\beta, \hat D_2, \cdots, \hat D_m)$ are considered. We also see that in the case $T = \hat\beta$, we have

$$n^{1/2}\| P_\beta - \Pi Q_\beta \| \le \sqrt{p}(\langle I^{-1}, F(\cdot, I^{-1}, \cdot)\rangle)^{1/2} = \sqrt{p}(\textstyle\sum_{i,j,k,l} F_{ik,lj}(I^{-1})_{ij}(I^{-1})_{kl})^{1/2}$$

which reduces to the statistical curvature $|\gamma(\beta)|$ in absolute value in the case $p = 1$, whereas in the multivariate case this quantity equals $\sqrt{p\ \mathrm{trace}(IE)}$, where E is the $p \times p$ matrix termed the "Efron excess" by Reeds (1975).

6. Comparison of test statistics. Consider a hypothesis of the form H_0: $H\beta = h_0$, where H is a $q \times p$ matrix of rank $q \le p$ and $h_0 \in V_0$ a known point. The most interesting example of this kind is testing that a coordinate of β takes a fixed value, but any "smooth hypothesis" may be written in this way, if necessary by a reparametrization. Let $\tilde\beta$ be the maximum likelihood estimate under H_0, and let H' be the transpose of H. We shall consider the following three test statistics of the hypothesis H_0:

$$\begin{aligned} L &= 2 \textstyle\sum_{i=1}^n (\log f(X_i, \hat\beta) - \log f(X_i, \tilde\beta)) \\ W &= (H\hat\beta - h_0)'(H'\hat I^{-1}H)^{-1}(H\hat\beta - h_0) \\ W_c &= (H\hat\beta - h_0)'(H'\hat J^{-1}H)^{-1}(H\hat\beta - h_0). \end{aligned}$$

Here L is the likelihood ratio test statistic, and W and W_c are quadratic test statistics in $(H\hat\beta - h_0)$ normalized by different estimates of its variance. In particular, W is the Wald test statistic and W_c a modified Wald test with $\hat J^{-1}$ as variance estimates of $\hat\beta$ instead of $\hat I^{-1}$. The index c means "conditional", although $\hat J^{-1}$ is not in general the conditional variance of $\sqrt{n}(\hat\beta - \beta_0)$ given A. The following theorem confirms a conjecture by Efron and Hinkley (1978), that even conditionally W_c follows a chi-squared distribution with error term of order $O(n^{-1})$.

THEOREM 6.1. *Under Conditions* 7.1 *and the assumption that* 7.1 (vi) *holds for the restricted model* H_0, *we have the following expansions under* H_0, *i.e. if* $H\beta_0 = h_0$,

(6.1) $$P_{\beta_0}\{L \le t \mid A = a\} = \chi^2_{p-q}(t) + O(n^{-1})$$

(6.2) $$P_{\beta_0}\{W_c \le t \mid A = a\} = \chi^2_{p-q}(t) + O(n^{-1})$$

$$(6.3)\qquad P_{\beta_0}\{W \le t \mid A = a\} = \chi^2_{p-q}(t) + O(n^{-1/2})$$

uniformly in $t \ge 0$ for all a in $\{\| a \|^2 \le (2 + \alpha)\log n\}$, where χ^2_{p-q} is the chi-squared distribution function with $p - q$ degrees of freedom.

The statement concerning W is in a sense negative and stated for comparison only. The important point is that the error is not in general $O(n^{-1})$. Note that marginally all three test statistics are asymptotically chi-squared distributed with error $O(n^{-1})$; see Chandra and Ghosh (1979).

Although this result indicates that L and W_c behave more like conditional tests than W does, it says nothing about the (marginal) properties of the tests. A possibility would be to compare the (asymptotic) powers of the tests, but a uniform superiority of any of these could hardly be expected. If one takes the standpoint in accordance with Example 1.1 that L is theoretically preferable to W and W_c, then one could compare W and W_c by their performance relative to L. This leads to the following result.

THEOREM 6.2. *Under the conditions of Theorem 6.1 W_c is stochastically closer to L than W is, in the sense that for any continuous function h: $\mathbb{R} \to [0, \infty)$, $h(0) = 0$, $h(x_2) > h(x_1)$ if $0 < x_1 < x_2$ or $x_2 < x_1 < 0$, we have*

$$(6.4)\qquad P_{\beta_0}\{h(\sqrt{n}\,(W_c - L)) < h(\sqrt{n}\,(W - L))\} = \delta(h) + o(1)$$

with $\delta(h) \ge 1/2$, and $\delta(h) = 1/2$ if and only if $F = 0$, and hence $W - W_c = O(n^{-1})$ with probability $1 - O(n^{-1})$.

Note that the function h is included to show that the result holds in "any scale", rather than, e.g., in the absolute values $| W_c - L |$ and $| W - L |$.

Both of the theorems suggest that W_c should be preferred to W, whereas it is hard to see any reason for preferring W to W_c in general. Moreover in connection with numerical maximization of the log-likelihood, W_c is easily computed because $-\hat{J}$ is just the matrix of second derivatives at the maximum. The results are, however, only asymptotic, and in particular cases W may well be preferable.

7. Conditions and proofs.

CONDITIONS 7.1. Let $\beta_0 \in \operatorname{int}(B)$ be a fixed parameter value, then

(i) If $x \in \{x; f(x; \beta_0) > 0\}$, then $f(x; \beta)$ is 7 times continuously differentiable w.r.t. β in a neighbourhood of β_0.

(ii) $I(\beta_0)$ is nonsingular and 5 times continously differentiable in a neighbourhood of β_0.

(iii) $E_{\beta_0}\{\| D^j \log f(X; \beta_0) \|^7\} < \infty$, $1 \le j \le 7$.

(iv) $\exists \delta_0 > 0$:

$$E_{\beta_0}\{(\sup\{\| D^7 \log f(X; \beta) \|;\ \| \beta - \beta_0 \| \le \delta_0\})^7\} < \infty.$$

(v) For $n = 1$ the characteristic function of $U(D_1, \cdots, D_7)$ belongs to L_m for some $m \in \mathbb{N}$, where U is an affine function mapping the affine support of $(D_1, \cdots, D_7)$ bijectively onto a real space, such that $\mathrm{Var}_{\beta_0}\{U\}$ equals the identity and $E_{\beta_0}\{U\} = 0$.

(vi) For sufficiently large n the MLE $\hat{\beta}_n$ of β exists with P_{β_0} – probability one, and for all $c > 0$

$$P_{\beta_0}\{\| \sqrt{n}(\hat{\beta}_n - \beta_0) \|^2 > c \log n\} = o(n^{-5/2}).$$

(vii) Expectations with respect to P_{β_0} of all linear and bilinear functions of $D \log f(X; \beta_0)$, $D^2 \log f(X; \beta_0)$ and $D^3 \log f(X; \beta_0)$ may be differentiated by differentiation under the integral sign.

We have not tried to minimize the assumptions of each theorem; instead, since the purpose of this section is to outline the techniques, they are a compromise between the demand that they should be easily verifiable, and the desire to avoid too great technicalities. In particular in (vi), probability one could be replaced by probability $1 - o(n^{-5/2})$. It may seem somewhat odd that 5 times differentiability is considered in (ii), and that 7 derivatives of $\log f$ are considered in (iv). These high numbers, compared to the theorems in which only second-order expansions are considered, are first of all used to derive the higher order expansions of $(\hat{\beta}, A)$ and A needed to control the error term of the expansion of the conditional distribution. In the sequel we shall refer to the assumptions as (i)–(vii), and it should be clear from the proofs what the purpose of each assumption is. Before going on to these we shall state a lemma of some independent interest.

LEMMA 7.2. *Let P be a probability measure and Q a finite signed measure both dominated by a measure μ on some measurable space (E, S). Let $f = dP/d\mu$ and $g = dQ/d\mu$ denote the densities. If $Q(E) = 1$ and a set $A \in S$ exists, such that for some $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$*

$$\text{(a)} \qquad \sup\{|f(x) - g(x)|;\ x \in A\} \leq \varepsilon_1$$

$$\text{(b)} \qquad \int_{A^c} |g(x)|\ d\mu(x) \leq \varepsilon_2$$

then

$$\sup\{|P(B) - Q(B)|;\ B \in S\} \leq 2(\varepsilon_1\mu(A) + \varepsilon_2). \tag{7.1}$$

PROOF.

$$|P(B) - Q(B)| \leq |P(B \cap A) - Q(B \cap A)| + |P(B \cap A^c) - Q(B \cap A^c)|$$
$$\leq \varepsilon_1\mu(A) + 1 - P(A) + \varepsilon_2 \leq 2(\varepsilon_1\mu(A) + \varepsilon_2). \quad \square$$

We shall now proceed to comment on the proofs, avoiding details that may in essence be found elsewhere.

Expansion of the distribution of $(D_1, \cdots, D_7)$. By the conditions (iii) and (v) we may apply Theorem 19.2 of Bhattacharya and Rao (1976) to obtain an asymptotic expansion as $n \to \infty$ in powers of $n^{-1/2}$ of the density of $n^{1/2}U(D_1, \cdots, D_7)$, the error term being $o(n^{-5/2})$ uniformly over the whole set.

PROOF OF THEOREM 4.1. We shall use Theorem 3.2 of Skovgaard (1980a) to transform the local expansion of U to a local expansion of (Z, A). This theorem is stated in terms of distributions, but since it is proved by the use of local expansions, it may be applied here in modified form. The technique was first used by Bhattacharya and Ghosh (1978) to derive an expansion of the distribution of Z under similar, but more general, assumptions. In Theorem 4.1 only the second-order expansions are stated, but to prove Theorem 4.2 we need to establish the validity of a local Edgeworth expansion with error term $O(n^{-2-\delta})$ for some $\delta > 0$. To do this a Taylor-series expansion of the form

$$Z \sim A_1(D_1) + n^{-1/2}A_2(D_1, D_2) + \cdots + n^{-5/2}A_6(D_1, \cdots, D_6) + o(n^{-5/2})$$

uniformly in $\| U(D_1, \cdots, D_7) \|^2 \leq c \log n$, is required. This is constructed as in Bhattacharya and Ghosh (1978) using conditions (i), (iv) and (vi). A similar expansion is needed for A, and this is obtained by expanding around $\hat{\beta} = \beta_0$ using the expansion of Z and conditions (i), (ii) and (iv). The expansion of A is only needed up to an error of order $O(n^{-2-\delta})$. On transforming the expansion of U, the validity of local Edgeworth expansions of (Z, A) and A including the n^{-2} terms is established, the errors being $O(n^{-2-\delta})$. Condition (vii) is needed to compute the second-order expansions, whereas we need not actually compute the higher-order expansions.

There is a slight technical problem in computing the differential $DF^{-1/2}(\hat{\beta} - \beta_0)$ of $F^{-1/2}$ in the direction $\hat{\beta} - \beta_0$. Since

$$(F^{-1/2})'F^{-1/2} = F^{-1}$$

and

$$DF^{-1}(\hat{\beta} - \beta_0) = -F^{-1}(DF(\hat{\beta} - \beta_0))F^{-1}$$

we obtain by the product rule

$$(DF^{-1/2}(\hat{\beta} - \beta_0))'F^{-1/2} + (F^{-1/2})'DF^{-1/2}(\hat{\beta} - \beta_0) = -F^{-1}(DF(\hat{\beta} - \beta_0))F^{-1},$$

which turns out to be all that is needed. Note that the right-hand side is independent of which "square root" of F is used. Based on the Taylor-series expansions, the computations of the κ's and the second-order expansions are straightforward; see, e.g., Skovgaard (1980b).

PROOF OF THEOREM 4.2. The method used to prove this is essentially the one given in Michel (1980). (4.11) is obtained by dividing (4.7) by (4.8); the problem is to prove the validity. To do this we need the expansions of $g_n(z, a)$ and $h_n(a)$ with error terms $O(n^{-2-\delta})$ as constructed above. The ratio of these will, on expanding in powers in $n^{-1/2}$ and keeping only the first- and second-order terms, give the same result as the ratio of the second-order expansions. The point is now that if α in Theorem 4.2 is sufficiently small, then the relative error of the higher-order expansion of $h_n(a)$ within the set $\| a \|^2 \leq (2 + \alpha)\log n$ is $O(n^{-1-\varepsilon})$ for some $\varepsilon > 0$. On this set also the error of the higher order expansion of $g_n(z, a)$ is $O(n^{-1-\varepsilon})$, when divided by $h_n(a)$. The theorem then follows from Lemma 7.2.

PROOF OF THEOREM 5.1. The main computations leading to (5.1) and (5.2) are quite similar to those given by Fisher (1925) (or Amari, 1982b). For example, to prove (5.2), an expansion of the likelihood equation gives

$$0 \sim D_1 + D_2(\hat{\beta} - \beta_0) + \tfrac{1}{2}D_3(\hat{\beta} - \beta_0)^2 + O(n^{-1/2}),$$

which in turn leads to an expansion of the form

$$D_1 \sim I(\hat{\beta} - \beta_0) + F^{1/2}A(\hat{\beta} - \beta_0) + Q(\hat{\beta} - \beta_0)^2 + O(n^{-1/2}),$$

where $Q(\hat{\beta} - \beta_0)^2$ is some quadratic form in $\hat{\beta} - \beta_0$. This shows intuitively that the conditional variance of D_1 given $(\hat{\beta}, A)$ is $O(n^{-1})$, although there are still some technical problems left. These problems are essentially overcome by showing that in the calculations of variances of D_1, one may neglect a region of the form $\| D_1 \|^2 > cn \log n$, where c is some constant. In this way the problems with integration of the error term may be avoided. It should also be noted that the results of Section 4 do not suffice to prove this theorem, but the (higher order) expansions used to prove Theorem 4.2 may again be used to establish the results.

Expansions of L, W and W_c. In the proofs of Theorem 6.1 and Theorem 6.2 we shall confine ourselves to the case of a simple hypotheses, i.e. H_0: $\beta = \beta_0$, since the ideas of the proofs are the same in the more complicated setting. Note that we then have $W = \hat{I}((\hat{\beta} - \beta_0)^2)$ and $W_c = \hat{J}((\hat{\beta} - \beta_0)^2)$. The Taylor-series expansions to second order of L, W and W_c around $Z = 0$ can be expressed as

$$\begin{aligned} L &\sim (I + n^{-1/2}F^{1/2}(A))(Z^2) + n^{-1/2}(\chi_{12}(Z^3) + \tfrac{2}{3}\,\chi_{111}(Z^3)) \\ W &\sim I(Z^2) + n^{1/2}(2\chi_{12}(Z^3) + \chi_{111}(Z^3)) \\ W_c &\sim (I + n^{-1/2}F^{1/2}(A))(Z^2) + n^{-1/2}(2\chi_{12}(Z^3) + \chi_{111}(Z^3)), \end{aligned}$$

the error being $O(n^{-1})p(A, Z)$ with probability $1 - O(n^{-1})$ uniformly on each set of the form $\| (Z, A) \|^2 \leq c \log n$, where p is a polynomial independent of n. These expansions are the key to the proofs of the two theorems of Section 6. Notice that the quadratic terms in Z are the squared length of Z as measured by the inverse *conditional variance* (cf. (4.12)) in L and W_c, whereas the *unconditional variance* is used in W.

PROOF OF THEOREM 6.1. Using the expansions above, (6.1) and (6.2) follows from Theorem 1 of Chandra and Ghosh (1979); see their Remark 2.2. Their condition (2.2) is not exactly fulfilled, because it only holds in sets of "size" $O(\log n)$ instead of $O(\sqrt{n})$, but it makes no essential difference in the proof. (6.3) is obvious, and it is seen that since $n^{-1/2}F^{1/2}(A)$ is in general *not* $O(n^{-1})$, neither is the error in (6.3).

PROOF OF THEOREM 6.2. Consider the differences

$$D = W - L \sim n^{-1/2}(-F^{1/2}(A)(Z^2) + \chi_{12}(Z^3) + \tfrac{1}{3}\chi_{111}(Z^3))$$

$$D_c = W_c - L \sim n^{-1/2}(\chi_{12}(Z^3) + \tfrac{1}{3}\chi_{111}(Z^3))$$

both being of order $O(n^{-1/2})$. To a first approximation, Z and $F^{1/2}(A)$ are independent, normally distributed with means zero and variances $\mathrm{Var}\{Z\} \sim I^{-1}$, $\mathrm{Var}\{F^{1/2}(A)\} \sim F$. Thus, to order $n^{-1/2}$, the conditional distribution of $\sqrt{n}D$ given Z is normal with mean $\sqrt{n}D_c$ and variance $F(Z^4)$, while D_c is a function of Z. In this approximate distribution it is seen that the probability of $h(\sqrt{n}D)$ being greater than $h(\sqrt{n}D_c)$ is a least ½, since the probability of the event that this occurs with D and D_c of the same sign equals ½. Since the other part of the event $h(\sqrt{n}D) > h(\sqrt{n}D)$ has probability zero if and only if F is zero, and hence $W = W_c + O(n^{-1})$, the theorem follows.

REFERENCES

AMARI, S. (1982a). Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika* **69** 1–18.

AMARI, S. (1982b). Differential geometry of curved exponential families—curvature and information loss. *Ann. Statist.* **10** 357–385.

BARNDORFF-NIELSEN, O. (1980). Conditionality resolutions. *Biometrika* **67** 293–310.

BHATTACHARYA, R. N. and RAO, R. R. (1976). *Normal Approximation and Asymptotic Expansions.* Wiley, New York.

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

CHANDRA, T. K. and GHOSH, J. K. (1979). Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-squared variables. *Sankhya A* **41** 22–47.

COX, D. R. (1975). Discussion to Efron (1975). *Ann. Statist.* **3** 1211.

COX, D. R. (1980). Local ancillarity. *Biometrika* **67** 279–286.

EFRON, B. (1975). Defining the curvature of a statistical problem. *Ann. Statist.* **3** 1189–1242.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

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

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **122** 700–725.

HINKLEY, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

LECAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 129–156.

LECAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

LECAM, L. (1975). Discussion to Efron (1975). *Ann. Statist.* **3** 1223–1224.

MADSEN, L. T. (1979). The geometry of statistical models. A generalization of curvature. Research report 79/1, Statistical Research Unit, Copenhagen.

MICHEL, R. (1978). Higher order asymptotic sufficiency. *Sankhya A* **40** 76–84.

MICHEL, R. (1979). Asymptotic expansions for conditional distributions. *J. Multivariate Anal.* **9** 393–400.

PEERS, H. W. (1978). Second-order sufficiency and statistical invariants. *Biometrika* **65** 489–496.

PIERCE, D. A. (1975). Discussion to Efron (1975). *Ann. Statist.* **3** 1219–1221.

RAO, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 531–545.

REEDS, J. (1975). Discussion to Efron (1975). *Ann. Statist.* **3** 1234–1238.

SKOVGAARD, I. M. (1980a). Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scand. J. Statist.* **8** 207–217.

SKOVGAARD, I. M. (1980b). Edgeworth expansions of the distributions of maximum likelihood estimators in the general (non i.i.d.) case. *Scand. J. Statist.* **8** 227–236.

ROYAL VETERINARY AND AGRICULTURAL UNIVERSITY
DEPARTMENT OF MATHEMATICS AND STATISTICS
THORVALDSENSVEJ 40
DK 1871 COPENHAGEN V
DENMARK