Ib M. Skovgaard

A Second-Order Investigation of Asymptotic Ancillarity

Preprint
June
1981
7

Institute of Mathematical Statistics
University of Copenhagen

Ib M. Skovgaard

A SECOND-ORDER INVESTIGATION OF ASYMPTOTIC ANCILLARITY

Preprint No. 7

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

June 1981

ABSTRACT. The paper deals with approximate ancillarity as discussed by Efron & 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. 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 & 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.

Key words: ancillarity, deficiency, Edgeworth expansions, loss of information, maximum likelihood, observed Fisher information, second order asymptotics, Wald's test.

1. Introduction

The purpose of this paper is to investigate some properties related to the conditioning on asymptotic ancillaries as proposed by Efron & 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 may be noted, that first-order asymptotics fails 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 essentially 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 & Hinkley (1978), we shall not go too much into this discussion, but merely given an example, essentially based on Pierce (1975), illustrating the advantages of this approach. Let $(\bar{X},\bar{Y})$ be the avarage of n independent two-dimensional normal statistics, each with the identity matrix as covariance and with mean $f(\beta) \in \mathbb{R}^2$, where β is a real parameter, and f is some smooth function. For each β, let L_β denote the line through $f(\beta)$ orthogonal to the tangent at β. If $\hat{\beta}$ is the maximum likelihood estimator of β, the 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 $f(\beta)$ 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 usual 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 observations 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 (x_2,y_2), a similar displacement would only change the estimate to $\hat{\beta}_2$. This suggests, that the <u>intrinsic accuracy</u> of the estimate is 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. In more general cases similar considerations hold, but the geometrical picture is not equally obvious.

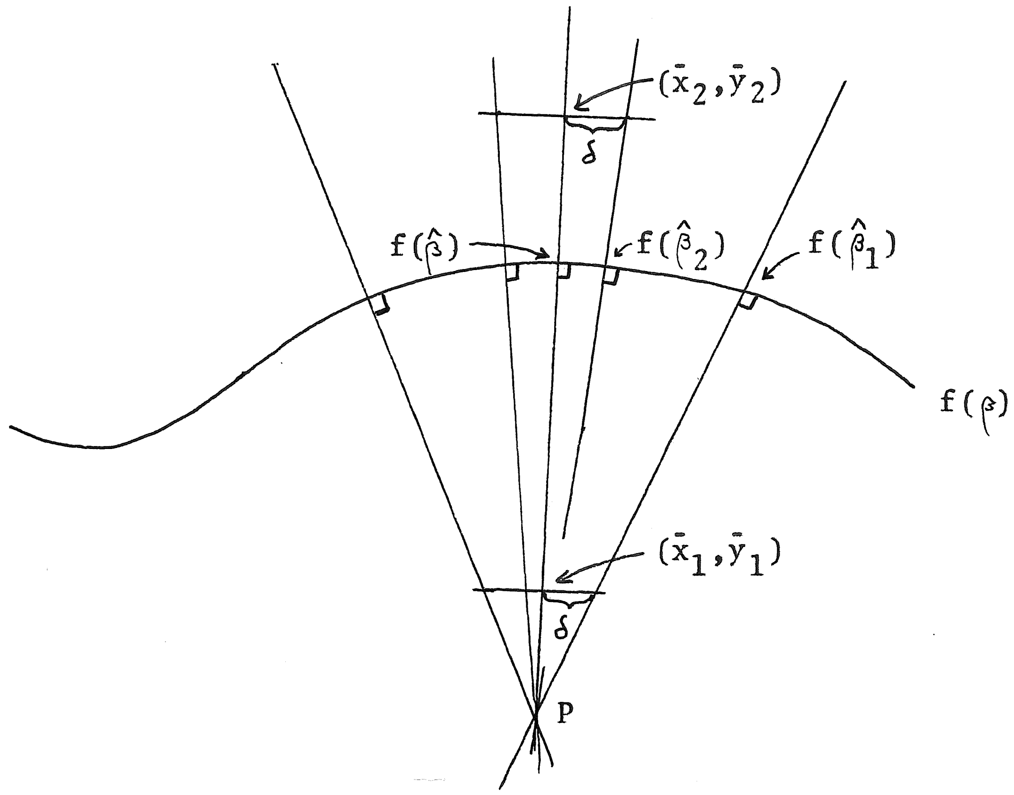

Fig. 1. Intrinsic accuracy of the maximum likelihood estimate. Sensitivity of the estimate due to a displacement δ of the observation depends on the distance of $(\bar{x},\bar{y})$ to P.

Several suggestions of 'ancillaries' capturing some of this additional information have been put forward (see 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 in more explicit form suggested by Efron & Hinkley (1978). A problem is, that this second derivative may contain a lot of information; however, after a suitable normalization, it will be asymptotically ancillary; more precisely it will be a locally second-order ancillary around the true parameter, see Cox (1980). Also, as will be shown in Section 4, the Fisher information contained in this statistic will tend to zero.

In Section 2 we provide the notation and the basic definitions. In Section 3 we derive the second-order Edgeworth expansion of the conditional distribution of the maximum likelihood estimator given the asymptotic ancillary statistic, and some moments of this distribution. In Section 4 we briefly investigate the loss of Fisher information in the various statistics and establish an implication of these results in terms of deficiencies as defined by LeCam (1964). Section 5 contains some comparisons of the Wald test statistics with the observed resp. the expected Fisher information as estimates of the inverse variance of the estimator. It is proved, that the conditional null-distribution of the former converges more rapidly towards its limiting chi-square distribution, and that this test statistic also has the advantage of being (marginally) stochastically closer to the likelihood ratio test statistic than the latter. Section 6 contains a simple example, and in Section 7 we state the regularity conditions and comment on the proofs.

Throughout the paper we are dealing with the i.i.d. case with multivariate parameter space. The results may be generalized to other cases along the lines of Skovgaard (1980a,b). Some important references concerning approximate ancillarity are Pierce (1975), Cox (1975, 1980), Efron & Hinkley (1978), Peers (1978), Barndorff-Nielsen (1980) and Hinkley (1980).

§ 2. Notation and setup

We shall use a coordinate free notation, which will allow us to handle the multivariate case almost as easily as the one - dimensional case. The reader, who does not wish to consider the problems arising in the multivariate case, will easily be able to recognize the meaning of the notation in the one - dimensional case without reading the first part of this section.

Let V_1 and V_2 be finite dimensional normed real vectorpaces and let $v, v_1 \in V_1$. By $B_j(V_1,V_2)$ we denote the vectorspace of j-linear symmetric mappings of $V_1^j = V_1 \times \ldots \times V_1$ into V_2, equipped with the norm

$$\| A \| = \sup\{ \| A(v^j) \| ; \| v \| = 1\} , A \in B_j(V_1,V_2)$$

where $v^j = (v,\ldots,v) \in V_1^j$. $C^s(B,V_2)$ denotes the s times continuously differentiable functions from an open set $B \subseteq V_1$ into V_2, and if $f \in C^s(B,V_2)$, then we define the k'th differential, $k \leq s$, of f at $v_1 \in B$ by

$$D^k f(v_1) \in B_k(V_1,V_2) , D^k f(v_1)(v^k) = \frac{d^k}{dh^k} f(v_1+hv)\Big|_{h=0}$$

where $h \in \mathbb{R}$. Also moments and cumulants of a random vector $Y \in V_1$ are regarded as multilinear forms, e.g. the k'th moment $\mu_k \in B_k(V_1^*, \mathbb{R})$ of y is given by

$$\mu_k(w^k) = E\{< w, Y >^k\}, w \in V_1^*$$

where V_1^* is the dual space to V_1, and $< , >$ denotes the inner product between a space and its dual. If V_1 is Euclidean, it is its own dual. An element $A \in B_k(V_1,V_2)$ may in a natural way be regarded as in $Hom(V_1,B_{k-1}(V_1,V_2))$ as the homomorphism given by $A(v)(v_1^{k-1}) = A(v,v_1^{k-1})$. We shall frequently use such natural constructions with the same notation for the two mappings. If $A: V_1 \times \ldots \times V_k \to V$, the we use the (matrix-like) convention, that a single argument refers to the last component of the domain (i.e. V_k), such that $A(v): V_1 \times \ldots \times V_{k-1} \to V$, $v \in V_k$.

Let $X_1, \ldots, X_n$ be independent identically distributed random variables on some measurable space, the distribution of X_i being a member of a family P_β, $\beta \in B \subseteq V$, where V is a Euclidean space of dimension $k \in \mathbb{N}$. We assume, that the family is dominated by μ, say, and let $f(x;\beta)$ denote some version of the densities. We also assume, that the conditions of § 7 are fulfilled, these essentially being various kinds of smoothness conditions.

Let $\beta_0 \in \text{int}\,(B)$ be the fixed true parameter value, and define

$$E_j(\beta) + S_j^{(n)}(\beta) = n^{-1} \sum_{i=1}^{n} D^j \log f\,(X_i;\beta) \in B_j(V,\mathbb{R})$$

such that $E_j(\beta)$ is non-random, and $S_j^{(n)}(\beta)$ has (P_β) expectation zero. Also, let

$$\chi_{i\ldots k}(\beta) = \text{cum}_\beta\,(S_i^{(1)}(\beta), \ldots, S_k^{(1)}(\beta))$$

denote the joint (P_β)- cumulant of $S_i^{(1)}(\beta), \ldots, S_k^{(1)}(\beta)$, i.e. a multilinear mapping of $B_i(V,\mathbb{R}) \times \ldots \times B_k(V,\mathbb{R})$ into $\mathbb{R}$. Finally we define (omitting the argument β)

$$I = -E_2 = \chi_{11}\,, \quad I^{(n)} = -\frac{1}{n} \sum_{i=1}^{n} D^2 \log f(X_i;\beta)$$

and $$F \in B_2(B_2(V,\mathbb{R}),\ \mathbb{R}) \quad \text{by}$$

$$F(a^2) = \chi_{22}(a^2) - I^{-1}(\chi_{12}(a)^2), \quad a \in B_2(V,\mathbb{R})$$

Notice, that $n\,\mathcal{I}$ is the 'expected' and $n\,I^{(n)}$ the 'observed' Fisher-information for the experiment, whereas F is the residual variance of $S_2^{(1)}$ after regression on $S_1^{(1)}$. In the sequel we shall use the convention, that if the argument β of a function is omitted, then the fixed argument β_0 is understood, and the value of a function at $\hat{\beta}_n$, the maximum likelihood estimator (MLE) of β, will be denoted by adding a circumflex, e.g. $\hat{I}^{(n)} = I^{(n)}(\hat{\beta}_n)$. Also, the index n will usually be omitted.

We are now ready to define the (hopefully) asymptotic ancillary statistic $A = A(X_1,\ldots,X_n)$ by

$$A = \sqrt{n}\;\hat{F}^{-\frac{1}{2}}(\hat{I}-\hat{\mathcal{I}}) \in W = B_2(V,\mathbb{R}) \tag{2.1}$$

which is just a normalized version of $\hat{I}$. Here $F^{-\frac{1}{2}} = \sqrt{F}^{-1}$, where $\sqrt{\ }$ is any smooth left square root, i.e. $\sqrt{F} \in \mathrm{Hom}(W,W)$ must satisfy $\sqrt{F}\,\sqrt{F}^{t} = F \in \mathrm{Hom}(W,W)$. It is immaterial, which inner product on W is used to identify W with its dual. It should be noted, that it is required (see §7), that F is regular. Otherwise a generalized inversed of $\sqrt{F}$ should be used as $F^{-\frac{1}{2}}$. This would not change the results, but notational difficulties would arise. Note, that $F^{-\frac{1}{2}}$ is not required to be symmetric, but is any linear variance normalizing transformation, when F is the variance.

§ 3. Expansion of the conditional distribution

In this section we shall expand the conditional distribution of $Z = \sqrt{n}(\hat{\beta}_n - \beta_0)$ given A under the distribution P_{β_0}. It is not hard to prove, that Z and A are asymptotically independent. Thus to obtain any interesting results, we must carry the expansion to second order, i.e. including the $n^{-\frac{1}{2}}$ terms.

The first step is to expand the simultaneous distribution of (Z,A). This is done in the following three steps. (i) Compute the second order (stochastic) Taylor-series expansion in terms of $S_1, S_2, \ldots$ around $0,0,\ldots$. (ii) Compute the first three joint cumulants of these approximating polynomials. These will be functions of the E_j's and the χ's. (iii) Using these cumulants, write down the Edgeworth approximation to the joint distribution. Since the expansion obtained in this way is the basis of all our results, we shall state it in detail in the following theorem.

Theorem 3.1. Under Conditions 7.1 we have the following local expansion for any $c > 0$

$$\sup\{|\, g_n(z,a) - \gamma_n(z,a)\,| \,,\ \|(z,a)\| \leq c \log n\} = O(n^{-1}) \quad (3.1)$$

$$\sup\{|\, h_n(a) - \zeta_n(a)\,| \,;\ \|a\| \leq c \log n\} = O(n^{-1}) \quad (3.2)$$

where g_n and h_n are the densities of (Z,A) and A, and

$$\gamma_n(z,a) = (2\pi)^{-(k+d)/2}(\det I)^{\frac{1}{2}}\{\exp -\tfrac{1}{2}(I(z^2) + \|a\|^2)\}$$

$$(1 + \kappa_z(I(z)) + \kappa_a(a) + \tfrac{1}{6}\kappa_{zzz}(I(z)^3) + \tfrac{1}{6}\kappa_{aaa}(a^3)$$

$$+ \tfrac{1}{2}\kappa_{zza}(I(z)^2, a) - \tfrac{1}{2}< I, \kappa_{zzz}(I(z))> - \tfrac{1}{2}< I, \kappa_{zza}(a)>$$

$$- \tfrac{1}{2}< 1_W, \kappa_{aaa}(a)>) \quad (3.3)$$

$$\zeta_n(a) = (2\pi)^{-d/2} \exp\{-\tfrac{1}{2} \| a \|^2\}(1 + \kappa_a(a) + \tfrac{1}{6} \kappa_{aaa}(a^3) - \tfrac{1}{2} < 1_W, \kappa_{aaa}(a) >) \tag{3.4}$$

where $d = \dim(W)$ and the κ's are the formal cumulants of (Z,A) as computed to order $n^{-\frac{1}{2}}$ from the second order Taylor series expansion of (Z,A) in (S_1,S_2). In particular

$$\sqrt{n}\ \kappa_z(I(z)) = -\tfrac{1}{2} < I^{-1}, \chi_{111}(z) + \chi_{21}(z) >$$

$$\sqrt{n}\ \kappa_{zzz}(I(z)^3) = -\ (2\,\chi_{111}(z^3) + 3\,\chi_{12}(z^3)) \tag{3.5}$$

$$\sqrt{n}\ \kappa_{zza}(I(z)^2, a) = -\ F(z^2,(F^{-\frac{1}{2}})^t(a))$$

Remark 3.2. It turns out, that $\kappa_{zaa} = 0$; otherwise the term $\frac{1}{2}\kappa_{zaa}(I(z),a^2) - \frac{1}{2} < 1_W, \kappa_{aaz}(I(z)) >$ should have been included in (3.3).

Remark 3.3. It is seen from (3.4) combined with the fact, that $\kappa_a = O(n^{-\frac{1}{2}})$, $\kappa_{aaa} = O(n^{-1})$, that A is not, in general, asymptotically second-order ancillary in the sense that the second order approximation to its distribution can be chosen to be independent of β_0. A is, however, locally second-order ancillary in the sense of Cox (1980), and this is the property, that turns out to be important to avoid loss of information (see § 4).

Theorem 3.4. Under Conditions 7.1 we have the following expansion of the conditional distribution of $Z = \sqrt{n}(\hat{\beta} - \beta_0)$ given $A = a$,

$$P_{\beta_0}\{Z \in B \mid A = a\} = \int_B \eta_n(z \mid a)dz + O(n^{-1}) \tag{3.6}$$

uniformly over all Borel sets $B \subseteq V$ and $\| a \|^2 \leq (2+\alpha)\log n$, for some $\alpha > 0$, where

$$\eta_n(z \mid a) = (2\pi)^{-k/2}(\det(I + F^{\frac{1}{2}}(a)))^{\frac{1}{2}} \exp\{-\tfrac{1}{2}(I + F^{\frac{1}{2}}(a))(z^2)\}$$

$$(1 + \kappa_z(I(z)) + \frac{1}{6}\kappa_{zzz}(I(z)^3) - \tfrac{1}{2} < I, \kappa_{zzz}(I(z)) >) \qquad (3.7)$$

Remark 3.5. Although the expansion (3.7) is easily obtained by dividing (3.3) by (3.4), it should be noted, that Theorem 3.4 does not follow from Theorem 3.1.

Remark 3.6. It is important to note, that the event $\{\| A \|^2 \leq (2+\alpha) \log n\}$ has probability $1 - O(n^{-1})$, such that Theorem 3.4 together with (3.2) implies, that

$$P_{\beta_0}\{Z \in B\} = \int_{\| a \|^2 \leq (2+\alpha)\log n} \zeta_n(a) \int_B \eta_n(z \mid a)\, dz\, da + O(n^{-1}) \qquad (3.8)$$

A local expansion of the conditional density of Z given A holding uniformly only on a bounded set, would not suffice to prove (3.8), and in this set the result would be incomplete.

There is a couple of things to note 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, see Skovgaard (1980b), 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

$$\tilde{V}(Z \mid A = a)^{-1} = \mathcal{I} + n^{-\frac{1}{2}} F^{\frac{1}{2}}(a) \sim \hat{I} + \mathcal{I} - \hat{\mathcal{I}}, \tag{3.9}$$

where $\tilde{V}$ is the variance of the approximate distribution. Thus it is seen, that the error $\mathcal{I} - \hat{\mathcal{I}}$ in approximating $\tilde{V}^{-1}$ by $\hat{I}$ is the same as the error in the usual (unconditional) approximation $\hat{\mathcal{I}}$ of $V\{Z\}^{-1} \sim \mathcal{I}$. If, in particular, the information is constant, then we have the approximation

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

in accordance with the result in Efron & Hinkley (1978) concerning the translation model. In fact, all that is needed for this to hold is, that the derivative of $\mathcal{I}$ at β_0 vanishes.

§ 4. Recovery of information

Fishers 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 $n\, I(\beta_0) = \text{int}\ (X)$, say, where $X = (X_1,\ldots,X_n)$. In general we let $\inf(T)$ denote the Fisher information (at β_0) contained in the 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 $\inf_A = E\{\inf(T \mid A)\}$. The well-known identity $\inf(T) = \inf(X) - E\{V\{n\, S_1^{(n)} \mid T\}\}$, 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 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).

<u>Theorem 4.1</u> <u>Under Conditions 7.1 we have</u>

$$\inf(X) - \inf(\hat{\beta}) = \text{tr}(I^{-1}F) + O(n^{-\frac{1}{2}}) \tag{4.1}$$

$$\inf(X) - \inf(\hat{\beta},A) = O(n^{-\frac{1}{2}}) \tag{4.2}$$

$$\inf(A) = O(n^{-\frac{1}{2}}) \tag{4.3}$$

$$\inf_A(\hat{\beta}) = \inf(X) - O(n^{-\frac{1}{2}}) \tag{4.4}$$

<u>where</u> $\mathrm{tr}(I^{-1}F) \in B_2(V,\mathbb{R})$ <u>is given by</u> $(\mathrm{tr}(I^{-1}F))(v^2) = F(v,I^{-1},v)$, $v \in V$.

<u>Remark 4.2</u>. The coordinate version of $\mathrm{tr}(I^{-1}F)$ is $(\mathrm{tr}(I^{-1}F))_{ij} = \sum_k \sum_l F_{iklj}\, g^{kl}$, where $F = (F_{iklj})$ and $I^{-1} = (g^{kl})$ are the coordinate versions of F and I^{-1}.

Note, that (4.4) follows from (4.2) and (4.3), since $\inf_A(\hat{\beta}) = \inf(\hat{\beta},A) - \inf(A)$.

Formal proofs of (4.1) and (4.2) 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 § 7, but we shall not elaborate on this point.

If one does not believe, as Fisher seemed to do, that the (Fisher) information is an absolute measure of information, then it would be natural to look for other interpretations or implications of Theorem 4.1 and similar results; see LeCam (1975). A reasonable possibility would be to measure the information lost in the reduction from $X = (X_1,\ldots,X_n)$ to $T = T_n(X)$ by the <u>deficiency</u> the experiment $(Q_\beta, \beta \in B)$ with respect to the experiment $(P_\beta, \beta \in B)$, when Q_β is the distribution of T; see LeCam (1964).

In agreement with LeCam (1956) (see also Michel (1978)) we shall use the slightly different measure

$$\delta_K(T,X) = \inf_{\Pi} \sup_{\beta \in K} \tfrac{1}{2} \| P_\beta - \Pi Q_\beta \|$$

$$= \inf_{\Pi} \sup_{\beta \in K} \sup_{A} \{ | P_\beta(A) - (\Pi Q_\beta)(A) | \} , \quad K \subseteq B \qquad (4.5)$$

where Π varies over the class of Markov - kernels and A over all measurable sets. Except for minor technical differences 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, and in any case the measure tells how well any test based on X can be reconstructed from T by a randomisation.

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 $(\Pi\, Q_\beta)$- 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

$$\| P_\beta - \Pi\, Q_\beta \| \leq \sqrt{k}\,(\mathrm{tr}\, R_\beta(T))^{\frac{1}{2}}\,(1 + o(1)) \tag{4.6}$$

where k = dim V 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 (4.6) then goes as follows

$$\| P_\beta - \Pi\, Q_\beta \| = \iint | f^t(x;\beta) - f^t(x;\hat{\beta}) |\, d\mu(x)\, d\, Q_\beta(t)$$

$$\sim \iint | (D_\beta \log f^t(x;\beta))(\hat{\beta}-\beta) |\, d\,\mu(x)\, d\, Q_\beta(t)$$

$$\leq \iint \| I(\beta)^{-\frac{1}{2}}(D_\beta \log f^t(x;\beta)) \|\, \| I(\beta)^{\frac{1}{2}}(\hat{\beta}-\beta) \|\, d\,\mu(x)\, d\, Q_\beta(t)$$

$$\leq (E_\beta\{(nI(\beta))(\hat{\beta}-\beta)^2\})^{\frac{1}{2}}(E_\beta\{ < (n\, I(\beta))^{-1},\ \inf_\beta(X \mid T) > \})^{\frac{1}{2}}$$

$$\leq \sqrt{k}\,(E_\beta\{\mathrm{tr}((n\, I(\beta))^{-1} \circ \inf_\beta(X \mid T))\})^{\frac{1}{2}}$$

$$= \sqrt{k}(\mathrm{tr}(\inf(X)^{-1}(\inf_\beta(X) - \inf_\beta(T))))^{\frac{1}{2}}$$

where the second inequality follows from Hölders inequality.

Using this result together with Theorem 4.1 we see, that $\delta_K(\hat{\beta},X) = O(n^{-\frac{1}{2}})$ and $\delta_K((\hat{\beta},A),X) = O(n^{-1})$, which has been proved more generally in Michel (1978). We also see that <u>in the case $T = \hat{\beta}$</u>, we have

$$n^{\frac{1}{2}} \| P_\beta - \Pi Q_\beta \| \leq \sqrt{k}(< I(\beta)^{-1}, \mathrm{tr}\, I(\beta)^{-1} F(\beta) >)^{\frac{1}{2}} \qquad (4.7)$$

$$(= \sqrt{k} \sum_{i,j,k,l} F_{ijkl}\, g^{il}\, g^{jk} \text{ in coordinates})$$

<u>which reduces to the curvature</u> $|\gamma(\beta)|$ in absolute value in the case $k=1$; see Efron (1975) for the definition and discussion of the curvature of a model.

§ 5. Comparison of test statistics

Consider a hypotheses of the form H_0: $H\beta = h_0$, where $H: V \to V_0$ is a known linear function 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. Let $\tilde{\beta}$ be the maximum likelihood estimate under H_0, and let $H^t: V_0^* \to V^*$ be the transpose of H. We shall consider the following three test statistics of the hypotheses H_0

$$L = 2 \sum_{i=1}^{n} (\log f(X_i, \hat{\beta}) - \log f(X_i, \tilde{\beta}))$$

$$W = (I^{-1} \circ (H^t, H^t))^{-1}((H\hat{\beta} - h_0)^2)$$

$$W_c = (\hat{I}^{-1} \circ (H^t, H^t))^{-1}((H\hat{\beta} - h_0)^2)$$

L is the likelihood ratio test statistic, and W and W_c are quadratic test statistic in $(H\hat{\beta}-h_0)$ mormalized with different estimates of its variance. W is the Wald test statistic and W_c a modified Wald test with $\hat{I}^{-1}$ as variance estimates of $\hat{\beta}$ instead of $\hat{I}^{-1}$. The index c means 'conditional', although $\hat{I}^{-1}$ is not in general the conditional variance of $\sqrt{n}(\hat{\beta}-\beta_0)$ given A. The following theorem confirmes a conjecture by Efron & Hinkley (1978); see also Cox (1980).

<u>Theorem 5.1</u> <u>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,</u> i.e. if $H\beta_0 = h_0$,

$$P_{\beta_0}\{L \leq t \mid A = a\} = \chi_p^2(t) + O(n^{-1}) \qquad (5.1)$$

$$P_{\beta_0}\{W_c \leq t \mid A = a\} = \chi_p^2(t) + O(n^{-1}) \qquad (5.2)$$

$$P_{\beta_0}\{W \leq t \mid A = a\} = \chi_p^2(t) + O(n^{-\frac{1}{2}}) \qquad (5.3)$$

uniformly in $t \geq 0$ for all a in $\{\|a\|^2 \leq (2+\alpha) \log n\}$, where χ_p^2 is the chi-square distribution function with $p = \text{rank}(H)$ degrees of freedom.

The statement concerning W is in a sense negative and stated for comparison only. The important thing is, that the error is not in general $O(n^{-1})$.

Note, that marginally all three test statistics are asymptotically chi-square distributed with error $O(n^{-1})$, see Chandra & Ghosh (1979).

Although this result indicates, that L and W_c behaves 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 the example in § 1, that L is theoretically preferable to W and W_c, although L is harder to compute, then one could compare W and W_c by their performance relative to L. This leads to the following result.

Theorem 5.2 Under the conditions of Theorem 5.1. W_c is stochastically closer to L, than W is, in the sense that for any continous function $h: \mathbb{R} \to [0,\infty[$, $h(0) = 0$, $h(x_2) > h(x_1)$ if $0 < x_1 < x_2$ or $x_2 < x_1 < 0$,

<u>we have</u>

$$P_{\beta_0}\{h(\sqrt{n}(W_c-L)) < h(\sqrt{n}(W-L))\} = \delta(h) + o(1) \qquad (5.4)$$

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

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. In particular cases there may, of course, be reasons for preferring W.

§ 6. An example

To illustrate some of the results we shall use the example at the end of the paper by Hinkley (1980). We shall not verify the Conditions 7.1, but it is a trivial matter apart from the regularity of F, which is not satisfied here. This causes however no problems.

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 compulations we get

$$\hat{\theta}_1 = \bar{Z} = \Sigma Z_i / n, \quad \hat{\theta}_2 = \Sigma Z_i Y_i / \Sigma Z_i^2$$

$$I(\theta) = \mathrm{diag}(1, 1+\theta_1^2), \quad \hat{I} = \mathrm{diag}(1, \Sigma Z_i^2 / n).$$

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

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

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 $Z_1, \ldots, Z_n$ 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

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

in agreement with (3.9). Also, as noted by Hinkley, $L = W_c = n(\bar{Z} - \theta_1)^2 + (\Sigma Z_i \varepsilon_i)^2 / \Sigma Z_i^2$ is exactly distributed as χ_2^2, whereas W deviates by an amount of order $n^{-\frac{1}{2}}$ (cf. Theorem 5.1 and 5.2).

§ 7. Conditions and proofs

Conditions 7.1. Let $\beta_0 \in$ int (B) be a fixed parameter value, then

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

(ii) $I(\beta_0)$ and $F(\beta_0)$ are regular, I is five and F four times continuously 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) The characteristic function of $U(S_1^{(1)},\ldots,S_7^{(1)})$ belongs to L_m for some $n \in \mathbb{N}$, where U is a linear function mapping the affine support of $(S_1^{(1)},\ldots,S_7^{(1)})$ bijectively onto a real space, such that $V_{\beta_0}\{U\}$ equals the identity.

(vi) For sufficently 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 the regularity of F is assumed for convenience only, and without this assumption the results would still hold with obvious modifications. In the same way in (vi) probability one could be replaced by probability $1-o(n^{-5/2})$. 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 $(E,\mathcal{A})$. Let $f = dP/d\mu$ and $g = dQ/d\mu$ denote the densities. If $Q(E) = 1$ and a set $A \in \mathcal{A}$ exists, such that for some $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$

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

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

then

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

Proof. $|P(B) - Q(B)| \leq |P(B\cap A) - Q(B\cap A)| + |P(B\cap A^c) - Q(B\cap A^c)|$

$\leq \varepsilon_1 \mu(A) + 1 - P(A) + \varepsilon_2 \leq 2(\varepsilon_1\mu(A) + \varepsilon_2)$. □

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

Expansion of the distribution of $(S_1,\ldots,S_7)$. By the conditions (iii) and (v) we may apply Theorem 19.2 of Bhattacharya & Rao (1976) to obtain an asymptotic expansion in powers of $n^{-\frac{1}{2}}$ of the density of $n^{\frac{1}{2}} U(S_1^{(n)},\ldots,S_7^{(n)})$, the error term being $o(n^{-5/2})$ uniformly over the whole set.

<u>Proof of Theorem 3.1</u>. 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 & Ghosh (1978) to derive an expansion of the distribution of Z under similar, but more general, assumptions. In Theorem 3.1 only the second-order expansions are stated, but to prove Theorem 3.4 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

$$Z \sim A_1(S_1) + n^{-\frac{1}{2}} A_2(S_1,S_2) +\ldots+ n^{-5/2} A_6(S_1,\ldots,S_6) + o(n^{-5/2})$$

uniformly in $\| U(S_1,\ldots,S_7) \|^2 \leq c \log n$, is required. This is constructed as in Bhattacharya & 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 compute the higher-order expansions.

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

Since

$$(F^{-\frac{1}{2}})^t F^{-\frac{1}{2}} = F^{-1}$$

and

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

we obtain by the product rule

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

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

Proof of Theorem 3.4. The method used to prove this is essentially the one given in Michel (1980). (3.7) is obtained by dividing (3.3) by (3.4); 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 $n^{-\frac{1}{2}}$ 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 3.4 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 4.1. We shall not comment on the main computations, which are quite similar to those in Fisher (1925), but only give a technical comment. Using the Edgeworth expansions, the variance of $S_1^{(n)}$ given $\hat{\beta}$ or $(\hat{\beta},A)$ is easily calculated, except that we need to show that a region $n\,\|S_1\|^2 > c\log n$ may be neglected. This follows however easily from the fact, that this is so in the marginal distribution of S_1.

<u>Expansions of L,W and W_c.</u> In the proofs of Theorem 4.1 and Theorem 4.2 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 $W_c = \hat{I}((\hat{\beta}-\beta_0)^2)$. The Taylor-series expansion to second order of L,W and W_c are obtained as

$$L \sim (I + n^{-\frac{1}{2}} F^{\frac{1}{2}}(A))(Z^2) + n^{-\frac{1}{2}}(\chi_{12}(Z^3) + \tfrac{2}{3}\,\chi_{111}(Z_1^3))$$

$$W \sim I(Z^2) + n^{-\frac{1}{2}}(2\chi_{12}(Z^3) + \chi_{111}(Z^3)) \qquad (7.2)$$

$$W_c \sim (I + n^{-\frac{1}{2}} F^{\frac{1}{2}}(A))(Z^2) + n^{-\frac{1}{2}}(2\chi_{12}(Z^3) + \chi_{111}(Z_1^3))$$

the error being $O(n^{-1})\,p\,(A,Z)$ with probability $1-O(n^{-1})$ uniformly on each set of the form $\|(Z,A)\|^2 \le c\log n$, where p is a polynomial independent of n. These expansions are the key to the proofs of the two theorems of Section 4. Notice, that the quadratic terms in Z are the squared length of Z as measured by the inverse <u>conditional variance</u> (cf. (3.9)) in L and W_c, whereas the <u>unconditional variance</u> is used in W.

<u>Proof of Theorem 5.1.</u> (5.1) and (5.2) follows from Theorem 1 of Chandra & Ghosh (1979), see their Remark 2.2. Their condition (2.2) is not exactly fulfilled, because it only holds in sets of 'size' $O(\log n)$ in stead of $O(\sqrt{n})$, but it makes no essential difference in the proof. (5.3) is obvious, and it is seen, that

since $n^{-\frac{1}{2}} F^{\frac{1}{2}}(A)$ is in general <u>not</u> $O(n^{-1})$, neither is the error in (5.3).

<u>Proof of Theorem 5.2.</u> Consider the differences

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

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

both being of order $O(n^{-\frac{1}{2}})$. To a first approximation Z and $F^{\frac{1}{2}}(A)$ are indpendent, normally distributed with means zero and variances $V\{Z\} \sim I^{-1}$, $V\{F^{\frac{1}{2}}(A)\} \sim F$. Thus, to order $n^{-\frac{1}{2}}$, 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 at least $\frac{1}{2}$, since the probability of the event, that this occurs with D and D_c of the same sign equals $\frac{1}{2}$. Since the other part of the event $\{h(\sqrt{n}\,D) > h(\sqrt{n}\,D_c)\}$ has probability zero if and only if F is zero and hence $W = W_c + O(n^{-1})$, the theorem follows.

References

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

Bhattacharya, R.N. & Rao, R.R. (1976). Normal Approximation and Asymptotic Expansions. Wiley, New York.

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

Chandra, T.K. & Ghosh, J.K. (1979). Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-square 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. & 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. on Math. Statist. and Prob. 1, 129-156.

LeCam, L. (1964). Sufficiency and approximate sufficiency. Ann. Math. Stat. 35, 1419-1455.

LeCam, L. (1975). Discussion to Efron (1975). Ann. Statist. 3, 1223-1224.

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). Ayymptotic efficiency and limiting information. Proc.Fourth Berkeley Symp. Math. Statist. and Prob. 1, 531-545.

Skovgaard, I.M. (1980a). Transformation of an Edgeworth expansion by a sequence of smooth functions. Preprint 1980, 2, Inst. Math. Stat., Univ. of Copenhagen.

Skovgaard, I.M. (1980b). Edgeworth expansions of the distributions of maximum likelihood estimators. Preprint 1980, 5, Inst. Math. Stat., Univ. of Copenhagen.

PREPRINTS 1980

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Olkin, Ingram and Væth, Michael: Maximum Likelihood Estimation in a Two-way Analysis of Variance with Correlated Errors in One Classification.

No. 2 Skovgaard, Ib M.: Transformation of an Edgeworth Expansion by a Sequence of Smooth Functions.

No. 3 Asmussen, Søren: Equilibrium Properties of the M/G/1 Queue.

No. 4 Johansen, Søren and Keiding, Susanne: A Family of Models for the Elimination of Substrate in the Liver.

No. 5 Skovgaard, Ib M.: Edgeworth Expansions of the Distributions of Maximum Likelihood Estimators.

No. 6 Tjur, Tue: A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model.

No. 7 Lauritzen, Steffen L. : Examples of Extreme Point Models in Statistics.

PREPRINTS 1981

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Johansen, Søren: Asymptotic Inference in Random Coefficient Regression Models.

No. 2 Asmussen, Søren: On the Role of a Certain Eigenvalue in Estimating the Growth Rate of a Branching Process.

No. 3 Lauritzen, Steffen L.: Time Series Analysis in 1880. A Discussion of Contributions made by T.N. Thiele.

No. 4 Asmussen, Søren: Conditioned Limit Theorems Relating a Random Walk to its Associate with Applications to Risk Reserve Processes and the GI/G/1 Queue.

No. 5 Johansen, Søren: The Statistical Analysis of a Markov Branching Process.

No. 6 Asmussen, Søren: Time - Dependent Approximations in some Queueing Sysetems with Imbedded Markov Chains Related to Random Walks.

No. 7 Skovgaard, Ib M.: A Second-Order Investigation of Asymptotic Ancillarity.