INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International
300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR

Information in the Presence

Of Nuisance Parameters

by

Bruce George Lindsay

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

1978

Approved by Ronald Pyke
(Chairperson of Supervisory Committee)

Program Authorized
to Offer Degree Biomathematics

Date August 4, 1978

UNIVERSITY OF WASHINGTON

Date: July 12, 1978

We have carefully read the dissertation entitled
Information in the Presence of Nuisance Parameters
submitted by
Bruce George Lindsay in partial fulfillment of
the requirements of the degree of Doctor of Philosophy
and recommend its acceptance. In support of this recommendation we present the following joint statement of evaluation to be filed with the dissertation.

Some of the most intriguing problems in statistical theory are those in which the nuisance parameters increase in direct proportion to the number of observations. Classical techniques of inference based on maximum likelihood often fail. Consequently resort is made to conditional, marginal, or other partial likelihood factorizations. However it has long been disputed whether such factorizations retain all the available information about the parameters of interest.

The present thesis resolves this issue for several statistical models which have important applications in epidemiology, genetics and other branches of scientific inquiry. By considering least favorable probability distributions on the space of nuisance parameters, the author introduces a new concept of statistical information which strengthens and generalizes the classical information measure of Sir Ronald Fisher. His new "K-information" leads to bounds for the variance of unbiased estimates which, for many problems of interest, are attained by statistics calculated from the partial likelihood. Sometimes no unbiased estimates exist in finite samples. For these and other situations a large sample theory is developed which, in parallel with the classical results for a fixed number of parameters, shows that K-information applies to consistent asymptotically normal (CAN) estimates, and that the lower bound may be achieved in the limit by statistics computed from the partial likelihood.

In deriving these important and fundamental results, the author demonstrates superb mastery of statistical theory and a very high degree of originality in his thinking. This thesis truly represents a major contribution to the modern theory of statistical inference.

DISSERTATION READING COMMITTEE: Norman Breslow

Arthur V. Peterson Jr.

Friedrich-Wilhelm Scholz

Doctoral Dissertation

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Bruce J Lindsay

Date August 4, 1978

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS

I wish to express my thanks and appreciation to all those who have contributed in a special way to this manuscript. Professor Norman Breslow provided the stimulation which led to the subject matter, plus assistance and supervision on the long road to the fruition of these ideas. Professor Art Peterson provided very helpful comments on the draft of this paper, too many of which had to be slighted due to the constraints of time. Professor Fritz Scholz contributed helpful suggestions in the formative stages. Special thanks are due to Professor Ronald Pyke. His stimulating classes and warm guidance were the primary sustenance of my education at the University. Finally, thanks are due to two people who provided moral support and love during these sometimes difficult years of graduate education: my wife Terry and my son Dylan.

CHAPTER I

Introduction, With a Review Of Efficiency

1.1 Chapter Introduction. This dissertation had its conception as an attempt to understand why (or in what way) certain statistical estimation techniques used in problems involving nuisance parameters might be "the best possible" instead of "the best available." The setting is parametric estimation in which there are two kinds of parameters: Those we are interested in estimating (hereafter called parameters of interest and denoted θ) and those which are necessary to model the problem correctly, but which are only a burden as far as estimation of θ goes (they will be called nuisance parameters and denoted ϕ). As will be demonstrated below, the presence of the nuisance parameters can distort maximum likelihood estimation, and so various techniques are used to find estimates which are less impacted by their presence. The following example will serve to fix ideas.

1.1.1 Example. Let X_{ij}, for $i = 1,2,\ldots,n$ and $j = 1,\ldots,k$, be independent and normally distributed with common variance θ and mean values $EX_{ij} = \phi_i$, which only depend on i. This is the classic one-way analysis of variance (ANOVA) model, with the parameter of interest being the common variance. The following set of statistics is sufficient:

$$\bar{X}_{i.} := k^{-1} \sum_{j=1}^{k} X_{ij} \quad , \quad i = 1,\ldots,n$$

$$S^2_{nk} := \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X}_{i.})^2 .$$

The maximum likelihood estimate (m.l.e.) of θ is

$$\hat{\theta}_{nk} := S^2_{nk} / nk.$$

If we hold n, which is the number of "treatments," fixed, and let k become infinite, then by standard m.l.e. results we have

$$\hat{\theta}_{nk} \xrightarrow{P} \theta \quad \text{as } k \to \infty.$$

If, however, we were to fix k, and let n, which is equal to the number of nuisance parameters, become infinite, then we have the following classic inconsistency result of Neyman and Scott (1948):

$$\hat{\theta}_{nk} \xrightarrow{P} \theta(k-1)/k \quad \text{as } n \to \infty.$$

As a replacement for the maximum likelihood estimate, one might use:

(1) $$s^2_{nk} := S^2_{nk} / n(k-1),$$

which is consistent as either n or k become infinite. As we will see later, the estimate (1) has several other natural derivations in addition to being a consistency-adjusted m.l.e. (i.e., MVUE, CMLE, Invariance). It is certainly the most commonly used estimate. We now ask, is this estimate in some way optimal in the asymptotic setting where n becomes infinite?

The investigation which led to the contents of this dissertation was in fact started because of a lower bound result by E.B. Anderson (1970, 1973) which suggested that it was possible that this estimate of the variance was not the "best possible" as n went to infinity. This dissertation demonstrates that the reason that classic efficiency results are not appropriate to this problem is that they fail to take into account the fact that the nuisance parameters remain essentially unknown as n becomes large. In particular, there are no consistent estimates of the ϕ_i's. A satisfactory way to account for their unknown character is to introduce into the key theorems concerning efficiency what are called mixing distributions over the nuisance parameter space. That is, in addition to considering densities $f(x;\theta,\phi)$, we also use mixed densities:

(2) $$f(x;\theta,P) := \int f(x;\theta,\phi)\, dP(\phi).$$

These densities are only used within the theorems. The actual initial modelling of the problem remains unchanged from the parameter space (θ,ϕ).

Although the idea of using mixed distributions (θ,P) seemed elegant and powerful in its infancy, it proved complex in its adolescence. As the reader will see, the chief problems in an asymptotic analysis (n goes to infinity) of a model such as Example 1.1.1 is that the nuisance parameter space becomes R^{∞}, infinite dimensional Euclidean space.

1.1.2 An Outline. The remainder of Chapter 1 is devoted to a review of classical efficiency results. It is intended as background material for the rest of the dissertation.

Chapter 2 states the problem of information and efficiency in the presence of nuisance parameters and briefly reviews procedures used to find good estimates. It attempts to answer the question: When is an estimate the "best possible?" Included is a brief discussion of the distinction between "best possible" and "admissible." Several key examples are presented, which, along with Example 1.1.1, are followed through the dissertation in order to elucidate and test the ideas.

Central to the techniques used to find good estimates in the presence of nuisance parameters is the reduction of the problem to one where the nuisance parameters play a smaller role. One such technique is discussed in Chapter 3, where estimators based on using m.l.e. methods on conditional likelihoods are developed. This chapter reviews the work of E.B. Andersen (1970, 1973) on conditional maximum likelihood estimators (CMLE). His extensive labors include showing the asymptotic normality of such estimators and deriving a lower bound theorem for estimation in this setting. It is in fact this bound that is referred to above as providing unsettling results in Example 1.1.1.

In Chapter 4, the first original material is introduced. The setting is lower bounds for unbiased estimation. We will see there that mixing distributions over the nuisance parameter space ϕ make a natural entry into the problem of information, and that the proper antecedents to the notion of using mixing distributions as a means to improve lower bounds were papers by E.W. Barankin (1949) and J. Kiefer (1951).

Two important concepts enter in this chapter. First, an information measure K is defined which is closely related to the Fisher Information I, but has the added power given by mixing distribution. Secondly, in the key

examples we find that the K-information in the problem is the same as the Fisher Information for the statistic we wish to use.

The fifth chapter takes an additional step to apply the notion of mixing distributions to lower bounds on the asymptotic variance of the asymptotic normal estimates. The way proves to be a bit tortuous, but it is shown that there is a legitimate asymptotic framework in which the lower bounds are more stringent than Andersen's. In particular, s^2_{nk} is shown to be optimal.

In the sixth chapter we turn to another concept of asymptotic efficiency based on Bahadur's notion (1971) of asymptotic effective variance. Despite its somewhat esoteric character, there are good reasons for its inclusion. The structure of the chapter is simpler than its predecessor, and the entry of the mixing distributions into the asymptotic structure come in a more natural manner. It is, in fact, in this setting that the significance of mixing distributions first became apparent.

We now turn to the second part of Chapter 1, a review of some of the classical information-efficiency results, based primarily on the treatment in C.R. Rao (1973). Although in the remaining chapters of the dissertation we will use a product parameter space $\Theta \times \Phi$, where Θ is open

in R^m, for the rest of this chapter there will be no nuisance parameters, and so the parameter space is Θ alone.

1.2 Fisher's Information. Suppose that X is a k-dimensional vector valued random variable, with a probability density $f(x;\theta)$ with respect to sigma-finite measure ν on R^k. For the moment assume m=1, so that θ is unidimensional. Assume that for any measurable set C in R^k

$$(3) \qquad \frac{\partial}{\partial\theta} \int_C f(x;\theta)\,d\nu = \int_C \frac{\partial}{\partial\theta} f(x;\theta)\,d\nu$$

This is satisfied under several different dominating conditions. (See M. Loeve {1977}, p. 127.) It is typically satisfied for exponential families. (See E.L. Lehmann {1959}, p. 52.) Define the score function to be the first derivative of the log likelihood. It will be denoted U or $U(\theta)$ depending on whether we need to emphasize its functional character. That is:

$$(4) \qquad U(\theta) := \partial \log f(X;\theta) / \partial\theta.$$

Notice that since $U = f'/f$, where the prime denoted differentiation with respect to θ, we have by equation (3) that

$$E(\ U(\theta)\ ;\ \theta) = \int f'(X;\theta)\, d\nu$$
$$= \partial \int f(x;\theta) d\nu/\partial\theta = \partial 1/\partial\theta = 0.$$

The θ following the semicolon in the expectation means that the expectation is taken with respect to density $f(x;\theta)$. Define the Fisher's Information in density f at θ to be:

$$I(\theta) := E(\ U(\theta)\ )^2 = \text{Var}\ U(\theta),$$

where E and Var are computed under θ.

For example, if we return to Example 1.1.1, and let $\phi_i=0$, $i = 1,\ldots,n$, then we have:

$$(5) \qquad U(\theta) = -\ nk\ /\ 2\theta + \sum_{ij} X_{ij}^2\ /\ 2\theta^2,$$
$$I(\theta) = nk/2\theta^2.$$

In the above setting, Fisher's Information has the following properties (see Rao {1973}, section 5a.4):

1.2.1 Additivity. Let I_1 and I_2 be the information in two independent random variables, and let I be the information in (X_1,X_2). Then $I = I_1 + I_2$. Hence if $X_1,\ldots,X_n$ are independent identically distributed random variables with information I in each X, their joint information is nI.

In Example 1.1.1, the information in each k-vector $X_i := (X_{i1}, X_{i2}, \ldots, X_{ik})$ is $k/2\theta^2$. When the means are all equal to zero, we have an i.i.d. problem, and additivity implies that the information in all n k-vectors is $nk/2\theta^2$.

1.2.2 Information in a statistic. Suppose T is a measurable function of X, with density $g(t;\theta)$ with respect to the measure ν, where the density g satisfies equation (3). Define the g-score function $U_g(\theta)$ to be the partial derivative of the log $g(t;\theta)$ with respect to θ, in direct analogy to (4). Define the Fisher information in T (or g) to be

$$I_g(\theta) := \text{Var } U_g(\theta). \tag{6}$$

We then have the following results:

a. $E\ (\ U(\theta)\ |\ T=t) = U_g(\theta)$.

b. $I \geq I_g$.

In our example, S^2_{nk} is distributed as θ times a chi-square with n(k-1) degrees of freedom. It follows that if g is the density of S^2_{nk}, then

$$I_g(\theta) = n(k-1)/2\theta^2 < I(\theta) = nk/2\theta^2.$$

1.2.3. Sufficiency. If $I_g = I$, then $U(\theta) = U_g(\theta)$ almost everywhere-ν, so

$$f(X;\theta) = g(T;\theta)\, H(X,T).$$

This is a situation in which T is sufficient for θ.

1.2.4 Definition. If θ is an m-dimensional vector, define the r-th score function as

$$U_r(\theta) := \partial \log f(X;\theta)/\partial\theta_r$$

Assume (3) holds for the partials $\partial/\partial\theta_i$. Let the Fisher's Information matrix I be that m by m matrix which has as its rs-th entry:

$$I_{rs} := E(U_r U_s).$$

Results analogous to 1.2.1 to 1.2.3 can be obtained. In particular, the matrix $I - I_g$ is non-negative definite.

If we compute the Fisher's information matrix for Example 1.1.1, where now we use $(\theta,\phi_1,\ldots,\phi_n)$ for $(\theta_1,\ldots,\theta_m)$, we get a diagonal matrix with diagonal entries:

(7) $$I_{11} = nk/2\theta^2 \;;\; I_{rr} = k/\theta \quad \text{for } r\neq 1.$$

1.3 Finite sample efficiency. (Rao {1973}, section 5c.2) The setting is as follows: we are trying to estimate $g(\theta)$, a real-valued function of θ, where θ is a real vector-valued parameter for a family of probability measures $\{F_\theta\}$ which can be expressed as densities $f(x;\theta)$ with respect to a sigma-finite measure ν. The observation vector is X. T is any unbiased estimate of g.

Fix a null parameter θ_0 through the rest of this section and let E(.) and Var(.) be understood to be taken at the null unless otherwise stated. The likelihood ratio

$$L(X;\theta) := f(X;\theta)/f(X;\theta_0)$$

is sometimes abbreviated to L or $L(\theta)$ in what follows. Now if F_θ is absolutely continuous with respect to F_{θ_0}, which we abbreviate to $\theta << \theta_0$, then L is finite a.e.-ν. Thus $L(X;\theta)\, f(X;\theta_0) = f(X;\theta)$ a.e.-ν and $E(L) = 1$. Also, if T is unbiased for g, we have

$$(8) \qquad E(T) = g(\theta_0) \quad ; \quad E(TL) = g(\theta).$$

It follows from the Cauchy-Schwartz inequality that:

$$(9) \qquad \text{Var } T \ \text{Var } L \geq (g(\theta) - g(\theta_0))^2.$$

This leads to the Chapman-Robbins lower bound result:

$$\text{(10)} \qquad \text{Var } T \geq \sup_{\theta} \{ (g(\theta)-g(\theta_0))^2 / \text{Var } L(\theta) \}$$

where the supremum is restriced to $\theta << \theta_0$. Notice that this result required no assumptions on the θ-space other than being an index space. If, however, Θ is an open subset of R, f has Fisher's information I, g is continuously differentiable, and the family of functions:

$$D(\theta) := (L(\theta) - 1)^2/(\theta - \theta_0)^2$$

is uniformly integrable for θ near θ_0, then (10) implies:

$$\text{Var } T \geq \lim_{\theta \to \theta_0} \frac{(g(\theta) - g(\theta_0))^2}{(\theta-\theta_0)^2} / D(\theta).$$

But $D(\theta) \to (f'/f)^2$ as $\theta \to \theta_0$, so

$$\text{(11)} \qquad \text{Var } T \geq (g'(\theta_0))^2/I(\theta_0).$$

The term to the right of the inequality in (11) will be referred to as the Cramer-Rao lower bound.

We can similarly derive a lower bound from the Chapman Robbins inequality (10) when $\theta = (\theta_1,\ldots,\theta_m)$ is vector-valued. Suppose $g(\theta)$ is a real-valued function with

continuous first partials in the θ-components, and set

$$(12) \qquad h_r(\theta) := \partial g(\theta)/\partial \theta_r \qquad r = 1,\ldots,m, \text{ and}$$

$$h := (h_1,\ldots,h_m)^t.$$

Then under regularity conditions on the density f, the multi-dimensional analogue of (11) is:

$$(13) \qquad \text{Var } T \geq h^t I^{-1} h$$

where I, the Fisher's information, is assumed to be non-singular.

Based upon the lower bound (13), we can define the Cramer-Rao efficiency of an unbiased estimator to be:

$$(14) \qquad \text{eff}(T) := h^t I^{-1} h/\text{Var } T$$

Notice that in the case when θ is unidimensional, the larger the information number I is, the smaller Var T has to be in order to achieve the same efficiency. This provides some justification for calling I an information. The larger it is, the better job of estimating we can theoretically do.

However, unless there is an estimate with efficiency 1, knowing the efficiency of an estimate does not tell if

it is the most efficient estimate. If for some g, the lower bound upon which the efficiency definition is based (in this case, Cramer-Rao) is met, it will be called tight. It is well known that the Cramer-Rao lower bound is not tight. In particular, it is known that only statistics T which are linear functions of the scores $U_1,\ldots,U_m$ can meet the bound. (Cox and Hinkley {1974}, p. 156.)

In Example 1.1.1 s^2_{nk}, defined in (1), is the minimum variance unbiased estimate of θ because it is a function of a complete and sufficient statistic (Lehmann-Scheffe Theorem). It has variance $2\theta^2/n(k-1)$, whereas the Cramer-Rao lower bound for $g(\theta) = \theta$ is $2\theta^2/nk$ from (7). Hence:

$$\text{eff}(s^2_{nk}) = (k-1)/k. \tag{15}$$

In Chapter 4, a lower bound for Example 1.1.1 is developed which is achieved by s^2_{nk}. Hence in that setting, s^2_{nk} will have efficiency 1.

1.4 Asymptotic efficiency. We first establish an asymptotic model. We suppose we have i.i.d. observations X_1, $X_2,\ldots$ from density $f(x;\theta)$, where f has Fisher's information I. We let $x^{(n)} := (x_1,\ldots,x_n)$ and call a sequence of

measurable functions $T_1(x^{(1)})$, $T_2(x^{(2)})$,... an estimator.

1.4.1 CAN estimates. It is often reasonable (thanks to the Central Limit Theorem) to restrict attention to estimators T_n which are consistent and asymptotically normal (CAN). That is, there exists a positive asymptotic variance $v(\theta)$ such that:

$$n^{-\frac{1}{2}}(T_n - g(\theta))/v(\theta) \xrightarrow{L} N(0,1) \text{ as } n \to \infty.$$

We might ask if there is a lower bound for the asymptotic variance $v(\theta)$. While it was once thought true that the Cramer-Rao lower bound

$$v(\theta) \geq h^t I^{-1} h \tag{16}$$

held, an example of its failure was provided by J.L. Hodges (see LeCam, 1953): Suppose T_n is CAN for θ, with asymptotic variance $v(\theta)$. Consider the estimator defined by

$$\begin{aligned} T_n' &= a\,T_n && \text{if } |T_n| < n^{-\frac{1}{4}}, \\ &= T_n && \text{else,} \end{aligned}$$

where a is a constant. Then T_n' is CAN for θ, with variance $a^2 v(0)$ at 0 and $v(\theta)$ elsewhere. It follows that there

can be no lower bound to the asymptotic variance of a CAN estimator.

An estimator is called superefficient if for all parameter values the estimate is asymptotically normal with an asymptotic variance never exceeding and sometimes less than the Cramer-Rao lower bound. What are we to make of superefficient estimators? Cox and Hinkley (1974) suggest that superefficiency is not a "statistically important idea" for the following three reasons. First, LeCam (1953) has shown that the set of points of superefficiency is Lebesgue measure zero. Second, if we were to construct hypothesis tests from these estimators, we would get no improvement in performance. Finally, for any fixed n, the reduction in mean squared error for parameter points near to the point of superefficiency is balanced by an increase in mean squared error at points a moderate distance away.

Whether or not superefficiency is statistically important, it is a mathematical nuisance. For this reason, several alternative versions of efficiency have been developed.

1.4.2 Fisher efficiency. Suppose we have i.i.d. observations with Fisher's information I in each observation. Assume θ univariate. Let $I(T_k)$ represent the Fisher's

information in T_k, as defined in 1.2.2. Let $I(T_n)/nI$ be the efficiency of T_n for n finite, and let the limit of the ratio go as n goes to infinity be the asymptotic Fisher efficiency of T_n, if the limit exists.

This definition of asymptotic efficiency has liabilities also. A statistic T_n can be fully efficient even if it is not consistent--the measure $I(T_n)$, while giving information about the discriminatory power of T_n's distribution, does not tell us if T_n is itself a good estimator.

1.4.4 CUAN estimates. Another approach to the problem of superefficiency is to further restrict the class of CAN estimates to eliminate those which are superefficient. This approach is similar to the one used in Chapter 5. We restrict ourselves to θ- univariate and $g(\theta) = \theta$.

CUAN estimators are defined to be CAN estimators for which the approach to normality of $n^{-\frac{1}{2}}(T_n-\theta)$ is uniform in compact intervals of θ. The reasonableness of this restriction comes from the following: One of the advantages of considering CAN estimators is that inference can be drawn using the normal distribution. Given a CAN estimator, a natural way to test the hypothesis $\theta = \theta_0$ is to use as a critical region, provided V is continuous,

$$(17) \qquad n^{-\frac{1}{2}} \, |T_n-\theta| \, / \, v^{\frac{1}{2}}(T_n) > d(\alpha).$$

Based on (17), we may be tempted to construct a confidence interval for θ of the form

$$(18) \qquad T_n \pm d(\alpha)\, v^{\frac{1}{2}}(T_n) n^{-\frac{1}{2}}.$$

However, this procedure is not justified unless the CAN estimator is also CUAN. That is, CUAN estimates fit naturally into a hypothesis testing framework. Further, the uniformity restriction suffices to eliminate any points of superefficiency: For CUAN estimates of θ, the lower bound for asymptotic variance at all null points θ is $I^{-1}(\theta)$.

1.5 Second order efficiency. Now we have defined a class of estimates (CUAN) for which the Cramer-Rao bound is valid, given sufficient regularity of the densities. Further, given some additional assumptions, a technique is known for reaching that bound: maximum likelihood estimates.

Having reached this plateau, we discover a new problem--there may be other fully efficient estimators, and we may want a way of distinguishing between them. In particular, if we have a multinomial distribution on m cells, with a probability distribution parameterized by θ real, then we might estimate θ by maximum likelihood, minimum

chi-square, or modified minimum chi-square: all are efficient in the CUAN sense.

Since this dissertation will deal only with what is called first order efficiency, this section is intended merely to call the reader's attention to the fact that a more sophisticated treatment of efficiency ("second order") has been worked on, with Rao (1963) and Efron (1974) being sources of discussion.

1.6 Asymptotic Effective Variance. Another approach to the problem of efficiency in an asymptotic setting is an offspring of the development of the theory of the probability of large deviations. The theory of asymptotic effective variance is more elegant mathematically than that of CUAN estimation, but it is less compelling as an important measure of the effectiveness of an estimator. This section is a brief synopsis of material to be found in R.R. Bahadur's SIAM publication, *Some Limit Theorems in Statistics* (1971) and material in several related articles (1960, 1967).

In Chapter 6 we will draw on the theory of asymptotic effective variance as an additional verification of the utility of mixing distributions as a tool for proving lower bound theorems.

We start with an information measure introduced by

S. Kullback and R.A. Leibler (1951).

1.6.1 Kullback-Leibler Information. Let P and Q be probability measures on some measurable space. If Q is not absolutely continuous with respect to P, let $K(Q,P) := \infty$. If Q is absolutely continuous with respect to P, let L(x) be the Radon-Nikodyn derivative, so that: $dQ = L(x)\, dP$. We then define

$$K(Q,P) := E(\ln L(x); Q).$$

Notice that $K(Q,P) = E(L \ln L ; P)$

1.6.2 Theorem. K is well-defined. $0 \leq K \leq \infty$. $K = 0$ if and only if $P = Q$.

For the proof of this, see Bahadur (1971).

The number K is an index of the statistical distance between P and Q. The smaller K is, the more difficult it is to discriminate between P and Q, as will be seen in what follows.

K is intimately related to the Fisher information I when in a parametric setting. In particular, let Θ be open in R^m, and for each θ in Θ, suppose $f(x;\theta)$ is a den-

sity with respect to a sigma-finite measure ν. Fix a null θ_0, let $L(X;\theta)$ be the likelihood ratio $f(X;\theta)/f(X;\theta_0)$, and let $K(\theta,\theta_0)$ denote the information distance between the measures associated with θ and θ_0. That is,

(19) $$K(\theta,\theta_0) = E(\ L(\theta)\ \ln\ L(\theta)\ ;\ \theta_0).$$

1.63 Theorem. Under sufficient regularity on the densities (see Kullback, 1968), we have

$$K(\theta,\theta_0) \sim \tfrac{1}{2}(\theta-\theta_0)^t I(\theta_0)\ (\theta-\theta_0) \text{ as } |\theta-\theta_0| \to 0$$

where ~ means that the ratio of the two sides goes to one as $|\theta-\theta_0|$ goes to zero.

For a proof, see Kullback (1968).

In particular, this means that if θ is univariate, then the second derivative of $K(\theta,\theta_0)$ with respect to θ, when evaluated at θ_0, is $I(\theta_0)$. The following lemma is analogous to the additivity of Fisher's information (see 1.2.2).

1.6.4 Lemma (Additivity) If X and Y are independent random variables whose densities generate informations K_1 and K_2

respectively, then the information for (X,Y) is K_1+K_2.

Proof. If L_1 and L_2 are the likelihood ratios for X_1 and X_2 respectively, then by independence the joint likelihood ratio can be represented as L_1L_2. Hence the joint information is

$$E(\ln L_1L_2\ ;\ \theta) = E(\ln L_1\ ;\ \theta) + E(\ln L_2\ ;\ \theta)$$
$$= K_1 + K_2.$$

We now turn to an important lemma, called Stein's lemma by Bahadur (1971), which relates hypothesis testing to the Kullback-Leibler distance. Choose and fix θ_0 and θ_1 and consider testing the simple hypothesis that $\theta = \theta_0$ versus the simple alternative that $\theta = \theta_1$. Let β be given, with $0<\beta<1$, and consider only tests with fixed power β against θ_1. For each n, let $\alpha_n=\alpha_n(\beta)$ be the infinum of all available sizes under the stated power requrements, when the sample size is n. By the Neyman-Pearson lemma, it is known that the infinum is attained by a possibly randomized test.

1.6.6 Lemma. For each β,

$$n^{-1} \log \alpha_n(\beta) \to -K(\theta_1,\theta_0) \text{ as } n \to \infty$$

This is proved in Bahadur (1971). A generalized version is proved in Chapter 6.

In addition to having an intrinsic interest of its own, this lemma turns out to be a key tool in the estimation theory which we now develop.

Suppose that g is a real-valued functional defined on Θ. For each n, let $T_n=T_n(X^{(n)})$ be a measurable function which is to be thought of as a point estimate of g.

1.6.7 Definition. Define $T_n=T_n(\varepsilon,\theta)$, the effective standard deviation of T_n, by

$$(20) \qquad P(|T_n-g(\theta)| \geq \varepsilon) = P(N(0,1) \geq \varepsilon/T_n).$$

Notice that if T_n is exactly normally distributed, with mean $g(\theta)$, then T_n is the actual standard deviation for each n. T_n is consistent if and only if the left hand side of (20) goes to zero as n goes to infinity, hence if and only if T_n goes to zero.

If T_n is CAN with asymptotic variance $v(\theta)$; then, for any positive h,

$$(21) \qquad \lim_{n\to\infty} n\, T_n^2 (n^{-\frac{1}{2}}h,\theta) = v(\theta).$$

In (21), we let $\varepsilon = n^{-\frac{1}{2}}h$ go to zero as n goes to infinity. If instead we hold ε fixed, we get a different kind of limiting result which we now present. Assume that Θ is open in R^m, and g is a continuously differentiable function of θ, with a vector of partials h. Assume that there exists an m by m positive definite symmetric matrix $I(\theta_0)$ such that for the fixed null point

$$K(\theta,\theta_0) \sim \tfrac{1}{2}(\theta-\theta_0)^t I(\theta_0)(\theta-\theta_0) \text{ as } \theta \to \theta_0. \tag{22}$$

Recall that by Theorem 1.6.3, (22) holds with I being the Fisher's information, given sufficient regularity. However, (22) can hold even under failure of the differentiability of the density as a function of θ. The double exponential is an example, used by Bahadur (1971), in which (22) holds even though the Fisher's information is not defined.

Under the above assumptions, the following result holds. It should be compared with (13) and (16).

1.6.8 Theorem. If T_n is a consistent estimate of g, then

$$\liminf_{\varepsilon \to 0} \liminf_{n \to \infty} nT_n^2(\varepsilon,\theta) \geq h^t I^{-1}(\theta)h \tag{23}$$

for every θ.

This is proved in Bahadur (1971).

1.6.9 Definition. Call the left hand side of (23) the asymptotic effective variance (a.e.v.) of T_n.

In view of Theorem 1.6.8, Bahadur suggests that an estimator be called efficient in the sense of a.e.v. if it attains the lower bound. The reader is referred to Bahadur (1960, 1967) for partial results concerning the efficiency of the MLE. The following lemma provides a simple means of computing a.e.v.

1.6.10 Lemma. If $Z_1, Z_2, \ldots$ are i.i.d. random variables such that $E(\exp(tZ))$ exists in some t-neighborhood of zero, then the a.e.v. of $\overline{Z}_n = n^{-1}(Z_1+\ldots+Z_n)$ as an estimate of $E(Z)$ is $\mathrm{Var}(Z)$.

For a proof, see Bahadur (1960).

1.6.11 Example. In Example 1.1.1, if we hold n fixed, and let k become infinite, then we have an infinite i.i.d. sequence. If we use $(\theta, \phi_1, \ldots, \phi_n)$ for $(\theta_1, \ldots, \theta_m)$, and consider estimates of θ, we get lower bound $2\theta^2/n$. The estimate s^2_{nk} is distributed like the average of k-1 i.i.d. observations from θ times a chi-square with n degrees of

freedom which has been divided by n. Applying lemma 1.6.10, the a.e.v. of s_{nk}^2 is $2\theta^2/n$. Hence the lower bound is tight, and s_{nk}^2 is fully efficient (a.e.v.).

CHAPTER II

Nuisance Parameters

2.1 Chapter Introduction. The focus of the dissertation is on the following quite general situation. For the observation random variable X there is a parametric probability model whose probability measures can be expressed as a family of densities $f(x;\theta,\phi)$ with respect to a sigma-finite measure ν. The parameter pair (θ,ϕ) come from a Cartesian product parameter space $\Theta \times \Phi$. Θ will be an open subset of R (univariate) and Φ will be required to be a measurable space so that we may put probability measures on it. In the examples, Φ will be a subset of R^k, with the usual Borel field of measurable sets. The parameter to be estimated will be θ; the nuisance parameter will be ϕ.

We will be interested in determining how well we can estimate θ in the absence of knowledge of ϕ. First consider the simple situation where ϕ is real-vector valued, and the conditions for the Cramer-Rao lower bound hold (Chapter 1, {13}). In this setting, the partition of the parameter space into θ and ϕ yields a corresponding partition of the information matrix I into submatrices I_{11}, I_{12}, I_{21}, I_{22}, where, in particular, I_{11} is the 1 by 1 matrix (scalar) in the upper left corner. We let I^{11}, I^{12}, I^{21}, and I^{22} be the corresponding partition of the matrix I^{-1}, when I is nonsingular. In this setting, when estimating functions $g(\theta)$ of θ alone, the Cramer-Rao lower bound

is of the form:

$$(1) \qquad \text{Var } T \geq h^t I^{11} h = h^2 I^{11},$$

where h, now a scalar, is the derivative of g with respect to θ. Call $(I^{11})^{-1}$ the information about θ available in the presence of ϕ, as it determines the lower bound for estimates of θ alone. Denote this $I(\theta;\phi)$. Unless otherwise stated, it will be this available information which the letter "I" will represent through the rest of the dissertation.

A standard result from matrix algebra is that

$$(2) \qquad I(\theta;\phi) = I_{11} - I_{12}\ I_{22}^{-1}\ I_{21}.$$

Notice that if the nuisance parameter ϕ was a known constant instead of an unknown parameter, the information about θ would be I_{11}. Thus the term $I_{12}I_{22}^{-1}I_{21}$ in (2) is a correction factor for the unknown character of the nuisance parameter.

The next lemma gives a computational device for finding the available information when we have independent observations with different nuisance parameters.

2.1.1 Lemma. (Additivity) Suppose X and Y are indepedent

observations from densities $f(x;\theta,\phi_1)$ and $g(y;\theta,\phi_2)$ respectively. Suppose they have available information $I^X(\theta;\phi_1)$ and $I^Y(\theta;\phi_2)$ respectively. Then the available information in (X,Y) is equal to $I^X + I^Y$.

Proof. If the parameter ordering (θ,ϕ_1,ϕ_2) is used then the Fisher's information matrices for X and Y are of the form:

$$I^X = \begin{bmatrix} I^X_{11} & I^X_{12} & 0 \\ I^X_{21} & I^X_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad I^Y = \begin{bmatrix} I^Y_{11} & 0 & I^Y_{12} \\ 0 & 0 & 0 \\ I^Y_{12} & 0 & I^Y_{22} \end{bmatrix}.$$

Because of independence, by 1.2.1 we have

$$I^{(x,y)} = I^X + I^Y.$$

The formula (2) for inversion gives the (1,1) entry of the inverse of $I^{(x,y)}$ as:

$$I^X_{11} + I^Y_{11} - I^X_{12}\,(I^X_{22})^{-1} I^Y_{21} - I^Y_{12}\,(I^Y_{22})^{-1} I^Y_{21}$$

$$= I^X(\theta;\phi_1) + I^Y(\theta;\phi_2).$$

2.1.2 Example. We again call on Example 1.1.1. The Fisher's information matrix for this problem is given in

(7) of Chapter 1. When $n = 1$, $I(\theta;\phi_1) = k/2\theta^2$. If we have n independent k-vectors $X_1,\ldots,X_n$, with means $\phi_1,\ldots,\phi_n$, then, by the additivity lemma, the joint information is $I^{(n)}(\theta;\phi^{(n)}) = nk/2\theta^2$.

2.2 Problem parameters. We now discuss cases where due to the unusual parameterization of ϕ, the classic means of determining lower bounds and efficiency are inapplicable.

2.2.1 Infinitely many real parameters. Suppose now that $X_1,\ldots,X_n$ is a sequence of independent random vectors where X_i comes from density $f(x;\theta,\phi_i)$. Here θ is a single fixed real parameter, but $\phi_i \varepsilon \Phi$ can vary from observation to observation. The X's are i.i.d. only if the ϕ's are constant. As n becomes infinite, then the nuisance parameter becomes an infinite sequence $\phi_\infty := (\phi_1,\phi_2,\ldots)$. Of course, the Fisher information is defined for any n finite, but the question remains as to what the asymptotic information is. As will be discussed in Chapter 3, E.B. Anderson (1973) used the limit as n becomes infinite of $I^{(n)}(\theta;\phi^{(n)})n^{-1}$.

Hereafter this general model will be called a Type I model. The reader will recognize that Example 1.1.1 is such a model.

2.2.2 Probability measure parameters. The parametric model $f(X_i;\theta,\phi_i)$ is the same as in the preceding section (2.2.1) but now the ϕ_i are viewed as an i.i.d. sequence from some unknown probability measure P on Φ. From this point of view, the observations X_i are themselves i.i.d. random variables from mixed density $f(x;\theta,P)$, defined in (2), Chapter 1.

Here the nuisance parameter space is properly viewed as $\mathcal{P}$, the family of probability measures on Φ. In this case, one cannot differentiate with respect to the nuisance parameter, so there is no Fisher information. Hereafter, this model will be called a Type II model.

If we view the means ϕ_i in the one way ANOVA model as having come from an unknown distribution P, then it is a type II model.

2.2.3 Function parameters. This type of parameter is well illustrated by the following famous example. Suppose that for each i, the observation T_i comes from the following density:

$$f(\ t\ ;\theta,\phi(.);z_i) :=$$
$$\phi(t)\ \exp(\theta z_i)\ \exp(-\int_0^t \phi\,(s)e^{\theta z_i}\,ds),$$

where t is restricted to the positive real axis. This is

the proportional hazards model for survival data proposed by D.R. Cox (1972). Here the function $\phi(.)$, called the underlying hazard rate, serves the role of nuisance parameter. The z_i are observed covariates whose presence keeps us from having i.i.d. observations. The θ parameter is the regression parameter which is to be estimated.

While this model is of substantial intrinsic interest as an example of nuisance parameters interfering with information about the parameter of interest, we concentrate here on the difficulties involved in the Type I and II models. See B. Efron (1977) for a modified Fisher's information approach to efficiency in Cox's model.

2.3 The Key Models. These are the models which we will follow through the discussion of the problems involved in determining the information present in the presence of nuisance parameters. They were chosen for a variety of reasons, and will represent a diversity of characteristics:

2.3.1 Model A. This is Example 1.1.1. It is a must for inclusion because of its importance in the development of ideas concerning nuisance parameters, starting with Neyman and Scott's 1948 paper. It is in some ways a less practical example than those which follow because one usually thinks of the mean vector as being the parameter of interest

in the one way ANOVA.

2.3.2 Model B. For this example, the observation vector for each i is a pair (X_i,Y_i) of independent Bernoullis. The respective success parameters are:

$$p_i := P(X_i=1) = \exp\theta+\phi_i \;/\; \{1 + \exp\theta+\phi_i\},$$

$$q_i := P(Y_i=1) = \exp\phi_i \;/\; \{1 + \exp\phi_i\}.$$

It follows that the log odds ratio, $\log p_i(1-q_i)/q_i(1-p_i)$, is equal to θ for all pairs of observations. The parameter θ is typically thought of as being a treatment effect when this model is used. This model, called a "matched pairs" model, can be found in D.R. Cox (1970), along with an example of its use.

2.3.3 Model C. This is just an extension of Model B to the case where the pair (X_i,Y_i) are binomials instead of Bernoullis. The same success parameters are used, and θ is still the log odds ratio for all pairs. Each pair can be thought of as being a 2 by 2 table. We let r_i and s_i represent the sample size parameters of X_i and Y_i respectively.

This model is separated from Model B because of mathematical distinctions which will be made later. This

model is treated in section 5.3 of Cox (1970).

2.3.4 Model D. For each i, we observe $X_i = (X_{i1},\ldots,X_{ik})$, a vector of i.i.d. observations from the following density:

$$f(x;\theta,\phi_i) = \exp\{-(x-\phi_i)/\theta\}/\theta \quad \text{if } x \geq \phi_i,$$
$$= 0 \quad \text{else.}$$

Thus $X_{ij} = \theta Y_{ij} + \phi_i$, where the Y_{ij}'s are independent unit exponentials. This classic problem is used by E.L. Lehmann (1959) in problem 12 of chapter 5 as an example of conditional hypothesis testing. It will turn out to present certain difficulties in the chapters to come, which is, of course, a valuable contribution.

2.3.5 Model E. For each i we observe a pair of independent Bernoulli random variables (X_i,Y_i), each with probability of success $\frac{1}{2} + \theta\phi_i$, where ϕ_i is +1 or -1 and θ is in the interval $(0,\frac{1}{2})$. The following genetics problem yields this model.

Suppose that for a certain type of bacteria there are two identifiable subtypes, A_0 and A_1. When a bacteria divides, each of the two progeny is assumed to have a probability p of changing from the subtype of its ancestor and a probability of 1-p of staying the same subtype. The

two progeny are assumed to mutate independently. We observe the subtypes of the progeny and identify the outcome with a pair of Bernoullis (X,Y), each of which is 1 if subtype A_1 is observed and 0 for subtype A_0. The ancestral subtype is assumed unknown.

Now if we model the probabilities with parameters (p,ϕ), where $\phi = +1$ if the ancestor is A_1 and $\phi = -1$ if the ancestor is A_0, then the pair (p,ϕ) is unidentifiable, as it generates the same probability distribution as $(1-p,-\phi)$. However, if we assume p is less than $\frac{1}{2}$ (or restrict ourselves to estimating $\theta := |\frac{1}{2} - p|$) then the problem reduces to model E.

2.4 Failure of the MLE. The examples are now used to demonstrate that the MLE commonly fails in the Type I model as n becomes infinite. This is done to motivate the consideration of other types of estimation in these models.

2.4.1 Model A. As mentioned in Section 1.1.1, Neyman and Scott (1948) proved that the MLE was inconsistent, and that, in fact, it converges in probability to $\theta(k-1)/k$.

2.4.2 Model B. For the n matched pairs of Bernoullis (X_i,Y_i), let n_1 be the number of pairs with $X_i+Y_i = 1$ and n_2 be the number with $X_i+Y_i=2$. Then the MLE of θ is

(Andersen {1973}):

$$(3) \qquad \hat{\theta}_n = 2 \log(n_1 - x_\cdot + n_2) - 2 \log(x_\cdot - n_2).$$

Andersen shows that $\hat{\theta}_n \xrightarrow{P} 2\theta$ as $n \to \infty$.

2.4.3 Model D. For each k-vector X_i, the MLE of ϕ_i is the minimum of $X_{i1},\ldots,X_{ik}$. With these $\hat{\phi}_i$ substituted into the likelihood equations, the MLE of θ is

$$(4) \qquad \hat{\theta}_{nk} = \sum_i \sum_j (X_{ij} - \hat{\phi}_i) \,/\, nk.$$

Note that (Johnson and Kotz {1970}):

$$E(X_{ij} - \hat{\phi}_i) = \theta + \phi_i - \theta/k - \phi_i = \theta(k-1)/k.$$

Also note that the terms $k^{-1} \sum_j (X_{ij} - \hat{\phi}_i)$ are i.i.d. over i, so that the strong law of large numbers holds. Hence

$$\hat{\theta}_{nk} \xrightarrow{a.s.} (k-1)\theta/k \qquad \text{as } n \to \infty.$$

Once again the MLE is inconsistent.

2.4.4 Model E. By sufficiency, the problem reduces to observing $S_i := X_i + Y_i$, a binomial $(2, \frac{1}{2}+\theta\phi_i)$ random

variable. The likelihood for each possible observation is:

$$P(S_i = 0) = (\tfrac{1}{2} - \theta\phi_i)^2,$$

$$P(S_i = 1) = 2(\tfrac{1}{2} - \theta\phi_i)(\tfrac{1}{2} + \phi_i\theta) = 2(\tfrac{1}{4} - \theta^2),$$

$$P(S_i = 2) = (\tfrac{1}{2} + \phi_i\theta)^2.$$

Thus the MLE for ϕ_i is clearly

$$\hat{\phi}_i = -1 \quad \text{if } S_i = 0,$$

$$\hat{\phi}_i = +1 \quad \text{if } S_i = 2,$$

$$\hat{\phi}_i = \pm 1 \quad \text{if } S_i = 1.$$

Substituting these back into the full likelihood, and letting n_1 be the number of times (out of n) that S_i=1, we get:

$$f(\, s^{(n)};\, \theta, \phi^{(n)}) = (\tfrac{1}{2} + \theta)^{2n-n_1}(\tfrac{1}{2}-\theta)^{n_1}$$

Maximizing this with respect to θ gives the unique solution to the likelihood equation: $\hat{\theta}_n = \tfrac{1}{2} - \tfrac{1}{2}\, n_1/n$. Next note that n_1 is itself binomial with sample size n and success parameter $2(\tfrac{1}{4}-\theta^2)$. This gives the result that

$$(5) \qquad E(\hat{\theta}_n) = \tfrac{1}{4} + \theta^2.$$

It follows by the strong law of large numbers that $\hat{\theta}$ converges almost surely to the right hand term of (5). Again, the MLE is inconsistent, although here the degree of bias depends on θ, and, in fact, decreases as θ increases.

2.4.5 MLE's for Type II models. Kiefer and Wolfowitz (1956) have shown that there exist, under certain assumptions, consistent maximum likelihood type estimates of (θ,P) in Type II models (see section 2.2.2). Despite the attractiveness of this result, there is a substantial practical difficulty involved in maximizing over the space of probability measures on Φ, and so this result appears to be of little practical importance. However, it should be noted that if the nuisance sequence $\phi_1,\phi_2,\ldots$ can be modeled as having come from a smooth parametric family of measures, then joint maximum likelihood estimation of the full set of parameters may yield a consistent and computable estimate of θ.

2.5 Ancillarity. Having demonstrated the dangers of maximum likelihood estimation, we now probe into alternative approaches to the problem of nuisance parameters. The first approach we consider involves the use of conditional likelihoods. However, lurking behind the use of conditional

likelihoods is the concept of ancillarity. Here is a brief introduction.

To start the discussion of ancillarity, assume for the moment there is no nuisance parameter. Following Cox and Hinkley (1974), Chapter 2, suppose that there is a minimal sufficient statistic S, where dim (S) is greater than dim (θ). If we can write S = (T,C), where C has a marginal distribution which doesn't depend on θ, then C is called an ancillary statistic (T is sometimes referred to as being conditionally sufficient). If C is a maximal such statistic, then the conditionality principle dictates that conclusions about the parameter of interest are to be drawn as if C were fixed at its observed value c--the only randomness to be considered by the statistician is that of T given C=c.

The arguments to be given for such a principle can be quite convincing, as the following example demonstrates. Suppose that a random variable T is known to be normally distributed with unknown mean μ, and with variance 1 or 10,000 depending on whether an independent Bernoulli random variable C turns up 0 or 1. Then S = (T,C) is sufficient, and C is ancillary. Suppose we observe (t,0). Is it reasonable that an interpretation of the data should be affected by the fact that the variance could have 10,000, but was not? As stated in Cox and Hinkley, "We may think

of C as an indicator of which 'experiment' was actually performed to produce the data."

Even if we subscribe to such a principle, there are some immediate difficulties in applying it. In addition to the problem of how C should be found, there are situations where C is not unique. See Cox and Hinkley (1974), example 2.28.

To extend the concept of ancillarity to the nuisance parameter situation, suppose S = (T,C) is minimal sufficient for (θ,ϕ), where the distribution of C depends on ϕ but not on θ, and the conditional distribution of T given C=c depends on θ but not ϕ. If the densities exist, this means there is a factorization of the form:

$$(6) \qquad f(t,c;\theta,\phi) = p(\, t|c \,;\, \theta \,)\, r(\, c \,;\, \phi \,),$$

where p and r are respectively the conditional and marginal densities. In this case, we also call C ancillary for θ, and T conditionally sufficient for θ in the presence of ϕ.

Although the justification for doing the analysis conditionally on C=c again seems quite reasonable, there are two points relevant to what follows. First, it is clear that maximizing the likelihood f in (6) consists of separately maximizing p over θ and r over ϕ. Hence the MLE of θ depends only on the conditional density p. This means

the MLE is the same as the conditional MLE, and so, as will be seen in the next chapter, it is typically consistent.

Second, the Cramer-Rao lower bound for θ also does not depend on r, in the following sense. If we define $I_p := E\,(\,\partial \ln p/\partial\theta)^2$ and $I_r := E(\,\partial \ln r/\partial\phi)^2$, then the Fisher information matrix for f is diagonal, with I_p and I_r as entries (1,1) and (2,2). So the available information about θ is $I(\theta;\phi) = I_p$, the Fisher's information in p.

Thus this notion of ancillary seems to fit in with the notions of information from Chapter 1. The marginal likelihood of C has no information in the sense that it has no input into the available Fisher's information about θ. The Cramer-Rao lower bound is based on the information in the conditional density.

2.6 Conditional Likelihoods. The logical next step beyond ancillarity is to consider conditioning on a statistic C whose distribution depends on θ, in addition to ϕ, but only in some yet-to-be-defined weak way. In this section we consider the problem of finding C, and in the next, the problem of weak ancillarity.

The aim is to find a conditional likelihood p which does not depend functionally on ϕ. Once done, maximization

of the conditional likelihood with respect to θ is often straightforward, resulting in conditional MLE's (CMLE). The properties of these estimates will be discussed in the next chapter. In this section, we ask the question, "How do we find a conditioning statistic C such that p has an optimal amount of information about θ?"

Andersen (1973) shows how to proceed: Fix θ. Suppose that the statistic C is minimally sufficient for ϕ with θ fixed, for all θ, and that C does not depend functionally on θ. Then the conditional probabilities given C=c do not depend on ϕ. Since C is minimally sufficient, it is unambiguously the optimal such factorization.

This procedure does not always work. Often the minimally sufficient C is a function of θ or is not a reduction from the full data at all (so that the conditional likelihood is uninformative). However, it works on all but one of the key models.

2.6.1 Model A. (One way ANOVA) The sample mean vector $\bar{X}^{(n)} := (\bar{X}_1., \ldots, \bar{X}_n.)$ is minimal sufficient for $\phi^{(n)}$. Thus we consider the conditional distribution of the data given $\bar{X}^{(n)}$. Since S^2_{nk} is independent of the sample mean vector, and they are jointly sufficient, it follows that the conditional distribution desired is exactly the marginal distribution of S^2_{nk}.

The CMLE for this distribution, as shown in Andersen (1973), is s^2_{nk}, as defined in (1) of Chapter 1.

2.6.2 Models B and C. (Paired Binomials) The statistics X_i+Y_i are minimal sufficient for the ϕ_i. Conditioning on the X_i+Y_i is equivalent to arguing conditionally on the marginal totals of a 2 by 2 table. For further discussion, see Cox (1970).

2.6.3 Model D. (Exponential with unknown support) We first transform the data, using the theory of the spacings of exponential random variables (see Johnson and Kotz {1970}). For each i, the order statistics $X_i^{(1)}<X_i^{(2)}<\ldots<X_i^{(k)}$ are sufficient for θ and ϕ_i. Transform these to new variables,

$$Y_i = kX_i^{(1)}$$

$$Z_{ij} = (k-j+1)\,(X_i^{(j)} - X_i^{(j-1)}) \quad \text{for } j = 2,\ldots,k$$

Then $\{Y_i - k\phi_i,\ Z_{ij} : i = 1,\ldots,n\ ;\ j = 2,\ldots,k\}$ are distributed as independent, identically distributed exponentials with mean θ. The vector $(Y_i, Z_{i2},\ldots,Z_{ik})$ is sufficient for (θ,ϕ_i) for each i.

Given this transformation, Y_i is minimally sufficient for ϕ_i for each i. By independence, the conditional

distribution of the data given $(Y_1,\ldots,Y_n)$ is equivalent to the marginal distribution of the n vectors $(Z_{12},\ldots,Z_{1k})$, $\ldots,(Z_{n2},\ldots,Z_{nk})$. It follows that the CMLE of θ is just the mean of the $n(k-1)$ Z_{ij}'s.

2.6.4 Model E. (Genetics Bernoulli) There are, in fact, no conditional distributions which depend only on θ. One argument for this is as follows: Fix θ. The factorization theorem (Lehmann {1959}, p. 49) states that if a statistic $T = T(s)$ is sufficient for ϕ, then the density may be written in the form

$$f(s;\theta,\phi) = g(T(s),\theta,\phi)\, h(s,\theta) \tag{7}$$

for some functions g and h. If T is a genuine data reduction, then $T(s_1) = T(s_2)$ for some $s_1 \neq s_2$, and the factorization (7) implies that the ratio $f(s_1;\theta,\phi)/f(s_2;\theta,\phi)$ does not depend on ϕ. A simple check reveals that this is not true for any of the possible (s_1,s_2) pairs in this problem. Hence S is itself minimal sufficient for ϕ. Since the conditional distribution of S given S is uninformative, the conditional approach yields nothing in this model.

2.7 Ancillarity and Information. What has been lost by using the conditional likelihood? Consider again Model A, the one-way ANOVA.

2.7.1 Model A. Let n = 1 for the moment, as this will suffice to demonstrate the problem. The sufficient statistics $(T,C) := (S^2_{1k}, \bar{X}_{1.})$ are independent, so by the additivity of Fisher's information

$$I_{(C,T)} = I_C + I_T,$$

where

$$I_C = \begin{bmatrix} 1/2\theta^2 & 0 \\ 0 & k/\theta^2 \end{bmatrix} \text{ and } I_T = \begin{bmatrix} (k-1)/2\theta^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Because the distribution of C depends on θ, there is a Fisher information loss when only the distribution of T is used. In this case it is $1/2\theta^2$.

Is this loss real, or is the information about θ in the distribution of the mean C too confounded with the nuisance parameter to be used? One might argue the latter for the following reasons. First, the minimum variance unbiased estimator of θ depends only on T. Second, the I_T matrix yields a Cramer-Rao lower bound which is met by the MVUE, which suggests I_T might be more accurately representing the information in the problem. Finally, as will be shown in the next section, T is a natural invariant for the problem.

On the other hand, C. Stein (1964) has shown that no estimates which are functions of S^2_{nk} are admissible under

squared error loss. In fact, there are estimates which have uniformly better risk. The point to be made is that results based on lower bounds, such as in Chapter 1, are cruder than results such as Stein's. However, the lower bound results of Chapter 1 do indicate that, if we let $k \to \infty$, the Stein type estimate cannot have superior asymptotic effective variance, and if it is asymptotically normal it cannot have superior asymptotic variance. That is to say, the additional information to be garnered by using the Stein estimate is negligible for sufficiently large sample sizes.

The above remarks concerning asymptotic results pertain only to the case where k becomes infinite. It still could be the case that the Stein estimator is strictly superior to s^2_{nk} as n becomes infinite, by one of the two criteria, asymptotic variance or asymptotic effective variance. This question will be answered in Chapters 5 and 6.

1.7.2 Models B and C. (Matched pairs and paired binomials) Here again there is controversy concerning the information in the conditioning statistic, X_i+Y_i. The conditional distribution of a paired binomial with X_i+Y_i fixed is that of a 2 by 2 table with all the marginal totals fixed.

R.A. Fisher (1935) originally proposed this conditioning argument:

> To the many methods of treatment hitherto suggested for the 2 by 2 table the concept of ancillary information suggests this new one. Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary...

That is, Fisher puts the marginal totals in the category of ancillary statistics, even though their distribution depends on θ, the parameter of interest. (Details about the marginal distribution and its Fisher information can be found in Chapter 4.)

There have been several attempts to justify Fisher's apparently intuitive remark about ancillarity. D. Basu (1977), O. Barndorff-Nielsen (1973), and Sprott (1975) consider various ways of clearly defining when one should use a conditional argument. In particular, Barndorff-Nielsen and Sprott disagree on the ancillarity of the marginal totals. Sprott concluded that "on occasion. . .the marginal totals apparently can contain information," because they did not satisfy his ancillarity criterion. On the other hand, they met Bardorff-Nielsen's M-ancillarity criterion.

Plackett (1977) took a different approach. He showed that using traditional inference procedures, one could not

obtain useful estimates or tests from the marginal distribution alone. This result leaves open the possibility that the marginal distribution contributes information to the full likelihood by a synergistic effect.

In this dissertation, the approach is taken that the proper measure of information in a statistical model should relate to how well that information can be used to estimate the parameters. In Chapters 4, 5, and 6 it will be demonstrated that the conditional likelihoods for Models A, B, C, and D found in Section 2.6 do contain all the information in the sense that by using them one can reach the lower bounds of those chapters.

2.8 Invariance. The one technique developed so far for limiting the role of the nuisance parameters has been conditioning. Given a density, there is a natural approach to finding the conditional likelihood depending on θ alone which has the most information about θ. Suppose we now focus on the marginal likelihoods instead. How does one go about finding a marginal likelihood which depends only on θ and has the most information about θ? The answer is not obvious. However, there is a technique which often yields such a marginal likelihood.

In a 1963 paper G.A. Barnard defined the notion of a statistic which is marginally sufficient by invariance.

Suppose that there are sufficient statistics (S,C) for (θ,ϕ) such that there exists a group G of transformations on the parameter space that leave θ fixed, but not ϕ. Further, the corresponding group of transformations on the sample space is such that it leaves S fixed, but is completely transitive on C. That is, the action of G can be used to move any outcome C=c to any other C=c'. Then S is said to be marginally sufficient.

This technique turns out to be quite useful on our key models:

2.8.1 Model A. The location transformation g defined by

$$x_{ij} \overset{g}{\to} x_{ij} + g_i, \quad i = 1,\ldots,n. \text{ and } j = 1,\ldots,k,$$

where $g_1,\ldots,g_n$ is a set of real numbers, corresponds to the transformation

$$\phi_i \overset{g}{\to} \phi_i + g_i \ , \ \theta \overset{g}{\to} \theta,$$

on the parameter space. Under the action of g, the means $\overline{X}_{i.}$ are completely transitive, and S^2_{nk} is fixed. Hence S^2_{nk} is marginally sufficient.

2.8.2 Model B and C. There are no transformations leaving

θ fixed.

2.8.3 Model D. (Exponential) We use the same sample space transformation as in 2.8.1. It yields the same transformation of the parameter space. In the notation of 2.6.3, $Y_1,\ldots,Y_n$ are transitive under g, and the scaled spacings Z_{ij}, $j\geq 2$, are fixed. Hence the latter are marginally sufficient.

2.8.4 Model E. (Genetics Bernoulli) The observations are Bernoulli pairs $(X_1, Y_1),\ldots,(X_n,Y_n)$. Recall that the binomials $S_1,\ldots,S_n$ were sufficient for the problem. Define the statistic W_i to be the indicator for $\{S_i=1\}$. The W_i's are independent Bernoullis with success parameter $2(\frac{1}{4}-\theta^2)$. Then for each i the two element group G_i on the same sample space consisting of the identity and $(x,y) \to (1-x,1-y)$ yields a corresponding group on the parameter space consisting of the identity and $(\theta,\phi) \to (\theta,-\phi)$. The same group leaves W_i invariant. It is transitive on X_i. A group of transformations for the entire sequence $(X_1,Y_1),\ldots,(X_n,Y_n)$ can be constructed as the Cartesian product of the groups G_i. The statistic $W^{(n)} := (W_1,\ldots,W_n)$ is invariant, and $X^{(n)} := (X_1,\ldots,X_n)$ is transitive.

Thus, for Model E invariance yields a marginal likelihood depending only on θ. Recall that conditioning failed

(2.6.4). On the other hand, for the paired Binomials, conditioning worked (2.6.2) but invariance failed.

2.8.5 Information and invariance. Invariance arguments are clearly a useful tool for reducing one's attention to a marginal likelihood which depends only on the parameter of interest. The question remains, how much information was lost in the process? One can obviate this question by appealing to an invariance principle. That is, one could presume that using invariant statistics was a concern overriding such issues as admissibility. If, however, one asks if one is getting the best possible estimates from invariant statistics, then one is returned to the controversies surrounding ancillarity. Since we use only the information in the marginal likelihood of the invariant, we seek justification for not using the information in the conditional likelihood of the data given the invariant statistic. This conditional likelihood will often depend on the parameter of interest θ. In particular, in Model A, S^2_{nk} is the invariant, so the conditional likelihood of the sample means $\bar{X}^{(n)}$ given S^2_{nk}, which by independence is just the marginal likelihood of the sample means, is the unused portion of the full likelihood. It has a distribution depending on θ. Hence the same controversy discussed in 2.7 is relevant here.

2.9 Partial Likelihood. A common approach to answering questions concerning the information lost when using a conditional or marginal likelihood for estimation has been to appeal to statistical principles such as invariance and ancillarity. But suppose a more sophisticated estimation is used, and that instead of using a simple marginal or conditional likelihood for estimation, we use a more complicated factorization of the full likelihood. If the likelihood factor used is no longer recognizable as having come from a single statistic, then conditionality and invariance principles are outmoded, but the question of information remains.

In 1975, D.R. Cox proposed the concept of partial likelihood, gave some important examples of its usefullness, and outlined some large sample results for estimates based upon maximizing the partial likelihood.

Suppose the vector Y of observations is transformed into a sequence $(X_1,S_1,X_2,S_2,\ldots,X_n,Y_n)$, where the components may themselves be vectors. Denote by p_i the conditional likelihood of S_i given $X^{(i)} := (X_1,\ldots,X_i)$ and $S^{(i-1)} := (S_1,\ldots,S_{i-1})$. Let r_i be the conditional likelihood of X_i given $X^{(i-1)}$ and $S^{(i-1)}$. Provided densities exist, the full density can be written

$$(8) \qquad f(\, x^{(n)}, s^{(n)}\,) = \prod_{i=1}^{n} p_i(s_i \mid x^{(i)}, s^{(i-1)})\; r_i(x_i \mid x^{(i-1)}, s^{(i-1)})$$

Cox defined $p := \prod p_i$ to be the partial likelihood based on S in the sequence (X_j, S_j). If p is a function of parameter θ, but not ϕ, then maximization of p often yields a CAN estimate of θ. (Cox, 1975).

Relevant to the discussion of conditionality and invariance is Cox's comment: "Both marginal and conditional likelihoods are in a natural sense ordinary likelihoods for derived experiments, but the same is not true in general for partial likelihoods; this is because of the way the conditioning events change." The partial likelihood, while a useful tool for isolating information about the parameter of interest, is thus not necessarily derived from the usual statistical principles.

Among the open problems mentioned by Cox is the following: "To specify circumstances under which all or nearly all the relevant information is contained in the partial likelihood." Since marginal and conditional likelihoods are special cases of partial likelihoods, the answer to this question is relevant to the discussions of information in 2.7 and 2.8.5. This dissertation is a discussion of a general solution to the problem of information in the presence of nuisance parameters, and so is an answer to

Cox's question. But before developing the necessary theory, we turn, in Chapter 3, to the work of E.B. Andersen. It is relevant at this point because it presents his answer to the question of information in conditional likelihoods, based on the Fisher's information matrix. A modification of his approach will be used in the original material in Chapters 4 through 6.

CHAPTER III

Conditional Maximum Likelihood Estimates

3.1 Chapter Introduction. Consider the type I model introduced in 2.2.1. The basic approach suggested by Chapter 2 is that one search for a factorization of the density of X_i of the form

$$f(x;\theta,\phi) = p(x;\theta)\, r(x;\theta,\phi),$$

where either p or r is a marginal density for a statistic C, and the other is the corresponding conditional. Once done, a natural approach is to use maximum likelihood estimation techniques on the partial likelihood $p_n = \Pi p(x_i;\theta)$.

There is a fundamental distinction to be made between a partial likelihood which is a product of conditionals and one which is a product of marginals. If p is the marginal density for a statistic C(x), then $C_1, C_2, C_3, \ldots$ is an i.i.d. sequence. The reason is that the distribution of C depends only on θ. It follows that the theory of maximum likelihood estimation for i.i.d. random variables holds.

If, however, the partial likelihood is a product of conditionals, classic maximum likelihood theory does not apply. Define the partial score statistic for the observation X_i to be the m-vector $U^p(Xi)$ with the k-th component

(1) $$\partial \ln p(X_i;\theta) / \partial \theta_k \ , \ k = 1,2,\ldots,m$$

Define the partial information to be the m by m matrix

(2) $$I^p(\theta;\phi_i) := E(U^p(X_i)(U^p(X_i))^t).$$

Notice that since p does not depend functionally on ϕ, the partial information depends on ϕ_i only through the expectation. If p is a conditional likelihood, with conditioning statistic C, then the conditional distribution of $U^p(X_i)$, given $C_i=c_i$, does not depend on ϕ. However, since the distribution of C typically depends on ϕ, the unconditional distribution of the score often does also. The upshot of this is that adjustments must be made in the usual maximum likelihood theory to take into account the fact that the scores are not identically distributed and that the informations may differ for each observation.

E.B. Andersen, in a monograph titled *Conditional Inference and Models for Measuring* (1973), produced the necessary adjustments of maximum likelihood theory needed to treat properly partial likelihoods that are products of conditional likelihoods. He established results concerning the existence, consistency, and asymptotic normality of the conditional MLE's. Since his material is important background for the material presented in this

dissertation, this chapter presents a brief summary of it. Also presented here is Andersen's lower bound for asymptotically normal estimates of the type I model, analogous to those discussed in Chapter 1. An improved version will be found in Chapter 5.

3.2 Assumptions on the model. The setting is the same as for the Type I and Type II models, with some additional restrictions. $X_1, X_2, \ldots$ is a sequence of independent random variables with a Euclidean range space, where X_i is distributed according to one of a parametric family of probability measures $\{ P(.;\theta,\phi) \}$. The parameter space is a product space $\Theta \times \Phi$, where Θ is an open subset of R^m and Φ is a topological space. As before, θ is fixed for all observations X_i, but ϕ is allowed to vary. It is assumed that the measures $P(\, . \, ; \theta,\phi)$ are absolutely continuous with respect to each other.

The next example reveals that unless the sequence of nuisance parameters has some assumptions placed upon it, there may be no reasonable estimators whatsoever.

3.2.1 Example. The observations X_i are independently normally distributed with mean θ and variance ϕ_i. The parameter of interest here, in contrast to model A, is the mean. Assuming the sequence of variances $\phi_1, \phi_2, \ldots$ to be

known, the MLE of θ is

$$(3) \qquad \hat{\theta}_n = \sum_{i=1}^{n} X_i \phi_i^{-1} \Big/ \sum_{i=1}^{n} \phi_i^{-1}$$

$$\overset{L}{=} N(\theta, \sigma_n^2), \text{ where } \sigma_n^2 = \Big(\sum_{i=1}^{n} \phi_i^{-1} \Big)^{-1}$$

It is clear that the MLE is consistent if and only if σ_n^2 goes to zero as n becomes infinite, which will be true if $\sum_{i=1}^{n} \phi_i^{-1} \to \infty$. In particular, it is inconsistent for $\phi_n = n2$, $n = 1,2,\ldots$.

The conclusion to be drawn is that for some sequences of ϕ's, reasonable estimates may not exist. In order to restrict the possible sequences, Andersen uses two different assumptions.

3.2.2 Assumption. The range space Φ of the ϕ's is a compact topological space.

3.2.3 Assumption. The sequence $\phi_1, \phi_2, \ldots$ is a sequence of i.i.d. random variables from an unknown distribution.

Assumption 3.2.2 is a restriction on the Type I model. The second assumption is identical to 2.2.2, the Type II model. It does not control the range of the ϕ's, but does prevent such irregular behavior as the sequence

(0,1,2,3,...). Notice that both assumptions are sufficient to make the estimate in example 3.2.1 consistent.

3.3 Sufficiency. Recall from Section 2.6 that the object in conditional likelihood estimation was to find a statistic C such that the conditional distribution of X given C does not depend on the parameter ϕ. Anderson ({1973}, Section 2.4) discusses various forms of sufficiency which might be relevant to this problem. As discussed in 2.6, he concludes that the most useful formulation is as follows: fix θ. Suppose that the statistic C, which does not depend functionally on the parameters, is minimally sufficient for $\{ P(\cdot ;\theta,\phi) : \phi\varepsilon\Phi \}$, for every θ. Then the conditional probability distributions of X given C do not depend on θ and is the appropriate likelihood to use. However, the technique did not work for model E, the genetics Bernoulli, because the minimal sufficient statistic C for each θ was equivalent to all the data. We find another failure of this technique in the following example.

3.3.1 Example. The model is the same as 3.2.1, except that now X_i is a k-vector (instead of a 1-vector) of i.i.d. normal variables, with mean θ and variance ϕ_i. For any fixed θ and for each i, the statistic

$$\sum_{j=1}^{k} (X_{ij}-\theta)^2$$

is minimal sufficient for ϕ_i. Since this statistic e-pends on the parameter of interest θ, Andersen's conditional likelihood approach fails to identify a conditional likelihood.

3.4 Conditional MLE's. This section reviews Andersen's formal derivation of CMLE's. Let ν be an arbitrary but fixed member of $\{ P(\cdot ;\theta,\phi) \}$. By the absolute continuity assumption of Section 3.2, the distribution of X_i has a density $f(x;\theta,\phi_i)$ with respect to ν. Let ν^* be the probability measure induced from ν by a measurable function C, so that $\nu^* (B) = \nu(C\varepsilon B)$ for any measurable set B. Since ν^* is finite, and the probability distribution of C is dominated by ν^*, C_i has a density $r(c;\theta,\phi_i)$ with respect to ν^*. Since the range space of X_i is assumed to be a Euclidean space it is known (Lehmann {1959}, p. 43) that there exists a version of the conditional probability of X_i given C_i=c which is a true probability measure. If $\bar{\nu}_c$ is the conditional probability measure given C_i=c, then

(4) $$p(x \mid c;\theta) := f(x;\theta,\phi) / r(c;\theta,\phi)$$

is a density for the conditional probability measures of X_i given $C_i=c$ with respect to the measure $\bar{v}_c$. The left hand side of (4) is independent of ϕ because of the sufficiency of C.

Now the X_i's are independent, so the conditional density of $X^{(n)}:=(X_1,\ldots,X_n)$ given $C^{(n)}:=(C_1,\ldots,C_n)$ is

$$(5) \qquad p_n(\, x^{(n)} \mid c^{(n)};\theta) := \prod_{i=1}^{n} p(x_i \mid c_i \,;\, \theta\,).$$

Define as the conditional maximum likelihood estimator (CMLE) that set of values of θ which maximize (5). Andersen notes that since two minimal sufficient statistics can only differ on sets of measure 0 with respect to the distribution of $(X_1,\ldots,X_n)$, the CMLE is essentially uniquely determined, provided that equation (5) has a unique maximum.

Further, Andersen proves that if one starts with densities for X and C based on sigma-finite measures μ and μ^* respectively, the conditional likelihoods one would obtain would only differ by a constant from (5), and so the inference would be identical.

3.5 Consistency of the CMLE. We start with the Type I model. In order to prove that the CMLE is consistent Andersen ({1973}, Section 2.6) specifies a set of assump-

tions about the parametric family of probabilities. Fix a null point θ_0 at which consistency is to be checked. The first assumption ensures that there is information about θ in the conditional likelihood:

3.5.1 Assumption. (Positive conditional information) Assume there exists a set B in the range space of C such that $P(C \in B; \theta_0, \phi)$ is positive for all ϕ in Φ and for each θ in a neighborhood of θ_0 and for all $c \in B$, there exists a set $A_c(\theta)$ such that

a. $p(x \mid c; \theta_0) \neq p(x \mid c; \theta)$ for all $x \in A_c(\theta)$

b. $P(A_c(\theta) \mid C=c ; \theta_0) > 0.$

Assumption 3.5.1 requires that the conditional density not be constant in θ for some neighborhood of θ_0, for a set of C-values with positive measure. The next assumption specifies the CMLE as the solution to the partial likelihood equations.

3.5.2 Assumption. (Unique solution) Assume $\log p(x|c;\theta)$ is for all x and c a differentiable function of θ, and the conditional likelihood equations

$$(6) \qquad \sum_{i=1}^{n} U^p(X_i;\theta) = 0,$$

where U^p is the partial score, have for almost all vectors $c^{(n)}$ a unique solution $\hat{\theta}_n^*$ in Θ which also maximizes the conditional likelihood $p_n(x^{(n)} \mid c^{(n)};\theta)$.

3.5.3 Assumption. (Continuity) Assume that $E(\log p(x|c;\theta);\theta_0,\phi)$ and $\mathrm{Var}(\log p(x|c;\theta);\theta_0,\phi)$ are, for all θ in Θ_0, continuous functions of ϕ, where Θ_0, a subset of Θ, is an open neighborhood of the null point θ_0.

3.5.4 Theorem (Andersen {1973}, p. 45), (Type I consistency) Under assumptions 3.5.1 through 3.5.3, the CMLE $\hat{\theta}_n^*$ converges almost surely to the true parameter θ_0.

In the Type II model, where the ϕ's are assumed to be an i.i.d. sequence from null distribution P_0, Anderson (1973) used the following assumptions:

3.5.5 Assumption. As in 3.5.1, except that B is a set in the range space of C such that $P(C\varepsilon B;\theta_0,P_0)$ is positive.

3.5.6 Assumption. Same as 3.5.2.

3.5.7 Assumption. For θ in an open neighborhood of θ_0, $E(\log p(x|c;\theta)\ ;\ \theta_0\ ,\ P_0)$ exists.

3.5.8 Theorem. (Andersen {1973}, p. 51), (Type II consistency) Under Assumptions 3.5.5 through 3.5.7, the CMLE $\hat{\theta}^*_n$ converges almost surely to θ_0 under the true distribution (θ_0, P_0).

We have already indicated that in our key models for which a CMLE exists (A through D), it is consistent. We now discuss Model B in more detail.

3.5.9 Example. Recall that in Model B, the matched pairs model, the statistic $C_i = X_i + Y_i$ was minimal sufficient for ϕ_i under fixed θ. (2.6.2.) The conditional distribution of X_i given $C_i = 1$ is Bernoulli, with success parameter $\exp(\theta)/(1+\exp(\theta))$. If $C_i = 0$ or 2, then X_i must be 0 or 1 respectively, and so the conditional measure is 1 if x_i is consistent with c_i, and 0 if not. For any sequence $(c_1, \ldots, c_n)$ with exactly n_1 ones and n_2 twos, $(x. - n_2)$ is the number of pairs (x_i, y_i) equal to (1,0). Hence the conditional likelihood is:

$$P(X^{(n)} = x^{(n)} \mid C^{(n)} = c^{(n)}; \theta) =$$

$$\exp\left((x. - n_2) \right) / \left(1 + \exp(\theta) \right)^n;$$

provided that $x^{(n)}$ is consistent with $c^{(n)}$. Maximizing the likelihood gives the CMLE:

$$\hat{\theta}^*_n = \log (x. - n_2) - \log (n_1 - x. + n_2)$$

In 2.4.2 the MLE $\hat{\theta}_n$ for this problem was given. Comparing with Equation (3) of Chapter 2, we see that $\hat{\theta}^*_n = \hat{\theta}_n/2$, and so is consistent, because (Section 2.4.2), $\hat{\theta}_n \xrightarrow{p} 2\theta$.

3.6 Asymptotic normality. Andersen ({1973}, section 3.2) also establishes conditions under which consistent CMLE's are asymptotically normal. Let $L^p = \log p$. As before, U^p is the partial score vector. Define the m by m matrix V^p by its ij-th entry:

$$(7) \qquad V^p_{ij}(\theta;\phi) := \partial^2 L^p / \partial\theta_i \partial\theta_j .$$

That is, V^p is the matrix of second partials of L^p.

The following assumptions are used in order to ensure the asymptotic normality of the CMLE in the Type I model:

3.6.1 Assumption. (Regularity) The first, second, and third partials of L^p exist for all θ in Θ. For all ϕ in Φ

a. $E(U^p; \theta, \phi) = 0$,

b. $E(-V^p; \theta, \phi) = I^p$, the partial information,

and there exist positive integrable functions H_{jk} such that for all θ in Θ,

$$\partial^3 L^p / \partial\theta_i\partial\theta_j\partial\theta_k \leq H_{jk}(x) \text{ for all } i,j,k.$$

3.6.2 Assumption. (Continuity) The density $f(x;\theta,\phi)$ is continuous in ϕ for all x. For all i,j,k the functions $\mathrm{Var}(V_{ij};\theta_0,\phi)$, $I^p_{ij}(\theta_0,\phi)$, $E(H_{jk};\theta,\phi)$, and $\mathrm{Var}(H_{jk};\theta_0,\phi)$ are finite, continuous functions of ϕ in Φ. In addition, $I^p(\theta_0;\phi)$ is nonsingular for all ϕ in Φ.

Now define, for null parameter $(\theta_0,\phi_1,\phi_2,\ldots)$,

$$I^p_n(\theta_0;\phi^{(n)}) := \sum_{i=1}^{n} I^p(\theta_0;\phi_i).$$

Since I^p_n is a symmetric square matrix, there exists a matrix B_n such that $I^p_n = B_n B_n^t$. Any other solution to $I^p_n = B_n B_n^t$ is an orthogonal transformation of B_n. Orthogonal transformations do not affect, however, the results in the sequel, so one need not be concerned with the exact form of B_n.

3.6.3 Lemma. (Asymptotic normality of score) Under assumptions 3.6.1 and 3.6.2, the random variable

$$\sum_{i=1}^{n} U^p(X_i;\theta_0)\, B_n^{-1}$$

converges in law, under null parameter $(\theta_0, \phi_1, \phi_2, \ldots)$, to an m-dimensional standard normal.

3.6.4 Theorem (Anderson {1973}, p. 84), (Type I asymptotic normality) Under assumptions 3.5.1 through 3.5.3 and 3.6.1 through 3.6.2, the CMLE $\hat{\theta}_n^*$ is asymptotically normally distributed with mean θ_0 and variance matrix $(I_n^p)^{-1}$. That is, $(\hat{\theta}_n^* - \theta_0)\, B_n$ converges in law to an m-dimensional standard normal distribution under $(\theta_0, \phi_1, \phi_2, \ldots)$.

In order to achieve normality under the Type II model the assumptions needed are:

3.6.5 Assumption. Let Θ_0 be an open neighborhood of θ_0, the true value of θ, and define

$$(8) \qquad I^p(\theta_0; P_0) := E\,(\, U^p (U^p)^t;\ \theta_0, P_0).$$

Then it is assumed that:

$$E\,(U^p;\ \theta_0, P_0) = 0$$

and

$$E(-V;\ \theta_0, P_0) = I^p(\theta_0; P_0).$$

Further, assume that there exist positive integrable (under (θ_0, P_0)) functions H_{jk} such that for all θ in Θ_0

$$\partial^3 L^p / \partial\theta_i \partial\theta_j \partial\theta_k \leq H_{jk} \quad \text{for all } i,j,k.$$

Finally, assume that I^p is nonsingular.

3.6.6 Theorem. Under assumptions 3.5.5 through 3.5.7 and 3.6.5, the CMLE converges in law to an m-dimensional normal distribution with mean θ_0 and variance matrix $(I^p)^{-1}$.

3.7 Uniform convergence to normality. Recall from Chapter 1 the two basic approaches to deriving a Cramer-Rao type lower bound for asymptotic variances. In the first (Section 1.4.1), the result held only for almost all θ (Lebesgue measure). In the second, a uniformity condition on the asymptotic normality yielded a bound which held for all θ. As reviewed next, Andersen, taking a mixed approach to the problem, assumes a uniformity only with respect to the nuisance parameter and derives a result which holds for almost all θ (Lebesgue measure).

3.7.1 Definition. A random variable $Y_n = Y_n(x^{(n)})$ is said to converge ϕ-uniformly in distribution to a distribution function G(x) if for each continuity point x of G,

$$\sup \{ \mid P(Y_n \leq x;\theta,\phi^{(n)}) - G(x) \mid \} \to 0$$
$$\text{as } n \to \infty.$$

The supremum is taken over the sequences $(\phi_1,\phi_2,\ldots)$.

3.7.2 Theorem. (Uniform convergence for Type I) Under the assumptions of Theorem 3.6.4, $(\hat{\theta}^*_n - \theta_0)\, B_n$ converges ϕ-uniformly to a standard normal distribution.

3.8 Andersen's Lower Bound. In a Type I model, after n observations, the parameters in the model are $(\theta,\phi^{(n)})$. We now suppose that the nuisance parameter space Φ is an open subset of R^s, for a positive integer s. Further, suppose the density displays the necessary regularity in θ and ϕ so that the full Fisher's information matrix I_n exists and is positive definite. Then a multivariate version of the Cramer-Rao lower bound is that for any unbiased estimate T of the m-vector θ, the matrix Var T - $I_n(\theta;\phi^{(n)})$ is nonnegative definite, where $I_n(\theta;\phi^{(n)})$ is as defined in Equation (2) of Chapter 2 (Rao {1973}, p. 326).

Now the CMLE is not characterized by its unbiasedness but rather by its asymptotic normality. Andersen (1973) therefore extends LeCam's lower bound result (1953) to the setting of the Type I model. The result stated

below is that the available Fisher's information $I_n(\theta;\phi^{(n)})$ determines a lower bound for asymptotic variance except on a θ-set of Lebesgue measure zero. His proof follows a proof by R.R. Bahadur (1964).

3.8.1 Theorem. (Andersen {1973}, p. 97), (Lower Bound) Suppose T_n is asymptotically normally distributed with mean θ and variance matrix V_n. Suppose the convergence is ϕ-uniform in the sense of Definition 3.7.1. Then if nV_n and $n^{-1}I_n(\theta;\phi^{(n)})$ both tend to a finite limit, the limit of $nV_n - n^{-1}I_n$ is nonnegative definite.

3.8.2 Example. In Model A, the estimator s^2_{nk}, which is the CMLE, is CAN, and satisfied the uniformity condition. It has asymptotic variance $2\theta^2/(k-1)$. The Andersen lower bound is the inverse of

$$\lim_{n\to\infty} n^{-1} I_n(\theta;\phi^{(n)}) = \lim_{n\to\infty} n^{-1}\ nk/2\theta^2 = k/2\theta^2.$$

Thus the estimator s^2_{nk} fails to meet Andersen's lower bound. We ask, is this because there are better estimates? Andersen ({1973} , p. 95) states his answer: "Failure of the CMLE to attain the lower bound does not exclude the estimator from having the smallest variance among ϕ-uniform asymptotically normally distributed

estimators. Cases of inefficiency are, therefore, difficult to identify." That is, Andersen's bound is not tight. It may by that the estimator s^2_{nk} has the smallest possible asymptotic variance. Indeed, it will be so shown in Chapter 5.

In Section 3.6 of his monograph, Andersen (1973) discusses sufficient conditions for his lower bound to be achieved. The following criterion turns out to be useful.

3.8.3 Lemma. The asymptotic variance of the CMLE attains the lower bound if $\log f(x;\theta,\phi)$ is a differentiable function of θ and ϕ and if there exist functions $d_i(\theta,\phi)$ such that

$$\partial \log r(c;\theta,\phi) / \partial\theta_i = d_i(\theta,\phi)\, \partial \log r(c;\theta,\phi)/\partial\phi,$$

for $i = 1,\ldots,m$ and almost all values of c.

Andersen defines a form of ancillarity weaker than that discussed herein in Section 2.5. Recall that C was called ancillary there if it had a distribution depending only on ϕ.

3.8.4 Definition. A statistic $C = C(x)$ is called weakly ancillary for θ in the presence of ϕ if for any given values (θ_0,ϕ_0) of (θ,ϕ), there exists a differentiable function $\phi(\theta)$ of θ with $\phi(\theta_0) = \phi_0$ such that for all θ

$$(6) \qquad r(c;\theta_0,\phi_0) = r(c;\theta,\phi(\theta))$$

for almost all values of c.

3.8.5 Theorem. If C is weakly ancillary (and hence if C is ancillary) the CMLE attains the Andersen lower bound.

3.8.6 Theorem. If C is ancillary in the weak or strong sense then the direct and the conditional MLE coincide, whenever both are obtained as unique roots of the likelihood equation.

Finally, we come to some examples to which Andersen has applied his lower bound theorem (3.8.1). The following section is a summary from his 1970 paper.

3.9 Andersen's Applications. The models to be discussed are all pairwise comparisons. Consider n independent pairs of independent random variables (X_i,Y_i) with distributions given by

$$dF(x) = \exp((\theta+\phi)x + p(x)) / c(\theta+\phi)\, d\nu(x)$$

and

$$dF(y) = \exp(\phi y + p(y)) / c(\phi)\, d\nu(y).$$

The statistic $T = X+Y$ is minimal sufficient for ϕ when θ is fixed. The CMLE for θ is given as the unique solution to

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \partial \log d(t_i,\theta) / \partial\theta,$$

where, if $A_t := \{ x : T(x) = t \}$, then

$$d(t_i,\theta) := \int_{A_t} \exp(\theta x + p(x) + p(t-x))\, d\nu(x|t).$$

By specifying the functions c, p, and ν in four different ways, the following four models are derived:

3.9.1 Example. (Normal) Let ν be Legesgue measure on R, let $p(x) = -\frac{1}{2}x^2$, and $c(w) = (2\pi)^{-1}\exp(\frac{1}{2}w^2)$. Then the joint density of (X,Y) is

$$f(x,y) = \exp(-\tfrac{1}{2}(x-\theta-\phi)^2 - \tfrac{1}{2}(y-\phi)^2) / 2\pi.$$

The estimation problem is that of estimating the difference in mean between two sets of normally distributed observa-

vations having the same variance. The CMLE is $\bar{x} - \bar{y}$. It attains the Andersen lower bound for asymptotic variance.

3.9.2 Example. (Poisson) Let ν be counting measure on the positive integers, $p(x) = -\log x!$, and $c(w) = \exp(\exp(w))$. The estimation problem is that of estimating the ratio of the mean values of two series of independent Poisson-distributed observations when this ratio is assumed to be constant. The Andersen lower bound for asymptotic variance is attained by the CMLE $\log \bar{x} - \log \bar{y}$.

3.9.3 Example . (Bernoulli) If we let ν be counting measure on $\{0,1\}$, $p(x) = 0$, and $c(w) = 1 + \exp(w)$, then the model is the same as Model B. The CMLE (see 3.5.9) does not achieve the bound.

3.9.4 Example. (Exponential) Let ν be Lebesgue measure on $(0,\infty)$, $p(x) = 0$, and $c(w) = (-w)^{-1}$ for w in $(-\infty,0)$. The joint density of (X,Y) is

$$f(x,y) = (-\theta-\phi)\,(-\phi)\exp(\theta x + \phi(x+y))$$

for x and y positive. That is, X and Y are independent exponentials, with respective means $(\theta+\phi)^{-1}$ and ϕ^{-1}.

Again the lower bound for variance is not attained by the CMLE, which is the solution to

$$\sum_{i=1}^{n} x_i = -n/\theta + \sum_{i=1}^{n} t_i \exp(t_i\theta) \,/\, (\exp(t_i\theta) - 1).$$

Andersen concludes this group of examples with the comment: "Thus for the two first models the lower bound is attained, while it is not the case for two other models. In all four cases the CMLE is a reasonable estimate for θ, but for the binomial (3.9.3) and gamma (3.9.4) versions of the exponential family the estimate fails to be completely efficient in the same sense of least possible asymptotic variance. Although no estimates with more optimal asymptotic behaviour are known to the author the question remains open whether such estimates exist."

CHAPTER IV

Information In Unbiased Estimation

4.1 Chapter Introduction. The central problem which this paper addresses has now been introduced: What are reasonable optimality and efficiency criteria in the presence of nuisance parameters? In the one way ANOVA (Model A), there is a traditional estimate for the variance, namely s^2_{nk}, which seems to be a reasonable estimate in the presence of the nuisance parameters, but which does not meet Andersen's lower bound for asymptotic variance. In Models B and C, the paired binomials, there is the same discrepancy between the intuitive feeling that the CMLE might be the most desirable estimate and the theory of Andersen. In Model D, the exponentials with unknown support, there is a natural estimate which is UMVUE, CMLE, and invariant-based MLE, but since Fisher's information is undefined, we have no conclusions about asymptotic efficiency. In Model E, there is again no Fisher's information, but there does exist an invariant marginal which seems to be highly informative. Does it provide fully efficient estimators?

The question of the optimality of the preceding estimators in the presence of nuisance parameters is not answered any differently using Fisher efficiency or Cramer-Rao efficiency. However, there is a substantial body of knowledge about unbiased estimation which warns against

taking the Cramer-Rao lower bound too seriously. Not only do we know that the lower bound is not tight, but we also know better ways of determining if an unbiased estimate is best possible in terms of variance; for example, using the Lehmann-Scheffe theorem (see Bickel and Doksum {1977}, p. 122).

However, it is a statistician's expectation that everything will be fine if the sample is large enough since, in typical problems, the asymptotic bound based on Fisher's information will be tight and will be attained by maximum likelihood estimation. It is disturbing therefore to find an asymptotic model in which MLE's are often inconsistent and the Fisher information lower bound may not be tight.

The logical place to start to climb the asymptotic heights is in the lowlands of unbiased estimation. In this chapter, finite sample results will be shown to give substantial clues to the asymptotic efficiency questions. In Model A the estimator s^2_{nk} has an efficiency of (k-1)/k based both on the Cramer-Rao lower bound (see Section 1.3, equation (15)) and on the Andersen lower bound (see Section 3.8.2). It seems logical that if one could increase the Cramer-Rao lower bound by using a "more accurate" information measure than Fisher's I, so that $\text{eff}(s^2_{nk}) = 1$, then one might also be able to increase the asymptotic efficiency to 1 by using the limit (as $n \to \infty$) of the same information measure. This will, in fact, be the case.

4.2 J-Information. The exploration for better information measures starts with the Chapman-Robbins lower bound of Section 1.3, now interpreted in a nuisance parameter framework. The parameter θ is real-valued. The parameter ϕ is from an arbitrary indexing set. Thus the following discussion applies to the Type II model by the interpretation that ϕ is a probability measure. Fix a null point (θ_0,ϕ_0). If T is an unbiased estimate of g, then equation (10) of Chapter 1 becomes

$$(1) \qquad \text{Var } T \geq \sup_{(\theta,\phi)} \{ (g(\theta)-g(\theta_0))^2 \ / \ \text{Var } L(\theta,\phi) \},$$

where $L(\theta,\phi) := f(X;\theta,\phi)/f(X;\theta_0,\phi_0)$ and the supremum is taken over $(\theta,\phi)<<(\theta_0,\phi_0)$. Now recall that in the derivation of the Cramer-Rao bound (equation (11) of Chapter 1) the approach was to take the limit as $\theta\to\theta_0$ of the term within the brackets {} in equation (1) above. But now this term depends on ϕ in addition to θ.

This suggests the following approach. For each θ define

$$(2) \qquad J^*(\theta) := \inf_{\phi} \{ \text{Var } L(\theta,\phi) \},$$

where the infinum is restricted to ϕ such that $(\theta,\phi) << (\theta_0,\phi_0)$. If one thinks of Var $L(\theta,\phi)$ as a measure of the

distance between (θ_0,ϕ_0) and (θ,ϕ), then the distance J* gives an indication of how small that distance can be made by choosing "least favorable" ϕ's.

Using (2) in (1) gives the equation

$$(3) \qquad \text{Var } T \geq \sup_{\theta} \{ (g(\theta)-g(\theta_0))^2 / J^*(\theta) \}.$$

Now, if one defines $J(\theta_0;\phi_0) := \lim J^*(\theta)/(\theta-\theta_0)^2$, assuming the limit exists, then (3) leads to the following lower bound result:

$$(4) \qquad \text{Var } T \geq (g'(\theta_0))^2 / J(\theta_0;\phi_0).$$

This is a natural analogue to the Cramer-Rao lower bound, with J playing the role of I.

The same bound (4) may be derived without requiring the convergence of $J^*(\theta)/(\theta-\theta_0)^2$ by making the following definitions: Define the upper J-information to be

$$J^+ := \lim_{\theta \to \theta_0^+} \inf \; J^*(\theta) / (\theta-\theta_0)^2;$$

the lower J-information to be

$$J^- := \lim_{\theta \to \theta_0^-} \inf \; J^*(\theta) / (\theta-\theta_0)^2;$$

and the J-information to be

$$J := \inf \{ J^-, J^+ \}.$$

Then equation (4) holds. An upper or lower J-information will be called exact if "lim inf" can be replaced by "lim."

It is important to notice that in the derivation of the lower bound (4) there are no assumptions about differentiability of the densities with respect to ϕ. Hence it can be applied to Type I models D and E. Moreover, it will be applicable to the Type II models. This latter possibility will be postponed until Section 4.4.8.

4.2.1 Example. (Model D) We compute the J-information in a single observation $X := X_{ij}$ which is an exponential with mean θ and support (ϕ,∞). The information for the full model will be established later by an additivity lemma. Now J* is computed by minimizing Var $L(\theta,\phi)$, but since Var $L = E\,L^2 - 1$, attention may be restricted to the term $E\,L^2$ for the minimization.

Notice that $E\,L^2 = E(\,L^2\,;\,\theta_0,\phi_0\,) = E(\,L\,;\,\theta,\phi\,)$. So, if we let $E^*(\,.\,) = E(\,.\,;\,\theta,\phi)$, then

$$E\,L^2 = E^*\,L = E^*\left(\frac{\theta_0\,\exp(-x/\theta)\exp(\phi/\theta)}{\theta\,\exp(-x/\theta_0)\exp(\phi_0/\theta_0)}\right)$$

$$= (\theta_0/\theta)\,\exp\,(\phi-\phi_0)/\,\theta_0\;E^*\left(\exp\left\{\frac{x-\phi}{\theta}(-1+\theta/\theta_0)\right\}\right)$$

Since $(X-\phi)/\theta$ is a unit exponential under (θ,ϕ), and so has moment generating function $(1-t)^{-1}$ for $t<1$, the last

expression becomes

$$(5) \qquad = (\theta_0/\theta)\ \exp\{(\phi-\phi_0)/\theta_0)\}\ (2 - \theta/\theta_0)^{-1}$$

for $\theta<2\theta_0$. The requirement that $(\theta,\phi)<<(\theta_0,\phi_0)$ restricts the minimization of (5) to $\phi\geq\phi_0$. On this set, (5) is minimized at $\phi=\phi_0$, and so

$$(6) \qquad J^*(\theta) = (\theta_0/\theta)\ (2 - \theta/\theta_0)^{-1}-1$$

Now if J^* is twice differentiable at θ_0, with $J^*(\theta_0)=0$ and $J^{*\prime}(\theta_0) = 0$, then by L'Hopital's rule,

$$(7) \qquad \lim_{\theta\to\theta_0} J^*(\theta)/(\theta-\theta_0)^2 = J^{*\prime\prime}(\theta_0)/2$$

Applying this computational device to (6), one finds that

$$(8) \qquad J(\theta_0;\phi_0) = \theta_0^{-2}$$

Notice that because we used $\phi=\phi_0$ in the calculations, they were carried out as if ϕ were a constant rather than a parameter. That is, the calculations were restricted to the one parameter family of densities $f(x;\theta,\phi_0)$. Thus it is not surprising that the J-information in (8) is exactly the same as the Fisher's information in an exponential

random variable with mean θ_0.

4.2.2 Example. (Model E) We compute the information in a single Bernoulli random variable X with mean $\frac{1}{2}+\theta\phi$. The likelihood ratio for this problem is:

$$L(\theta,\phi) = \frac{(\frac{1}{2} + \theta\phi)^x(\frac{1}{2} - \theta\phi)^{1-x}}{(\frac{1}{2} + \theta_0\phi_0)^x(\frac{1}{2} - \theta_0\phi_0)^{1-x}}$$

Recall that ϕ was in $\{-1,+1\}$. Notice that $L(\theta,\phi_0) \to 1$, but $L(\theta,-\phi_0) \not\to 1$ as $\theta \to \theta_0$, provided $\theta_0 \neq 0$. It follows by Fatou's lemma that Var $L(\theta,-\phi_0) \not\to 0$ as $\theta \to \theta_0$. Since $L(x;\theta,\phi_0) \to 1$ as $\theta \to \theta_0$ for every x, and we are in a finite discrete sample space, Var $L(\theta,\phi_0) \to 0$ as $\theta \to \theta_0$. The conclusion is that for θ sufficiently close to θ_0,

$$(9) \qquad J^*(\theta) = \inf \{ \text{Var } L(\theta,\phi_0), \text{Var } L(\theta,-\phi_0) \}$$
$$= \text{Var } L(\theta,\phi_0)$$

Since once again the nuisance parameter can be considered as a constant, we may compute the Fisher's information for the Bernoulli random variable with mean $\frac{1}{2}+\theta\phi_0$ to get

$$(10) \qquad J(\theta_0;\phi_0) = (\tfrac{1}{4}-\theta_0^2)^{-1}$$

In Models D and E, the J-information essentially

ignored the nuisance parameter. One asks, is this reasonable? It is analogous to a diagonal Fisher's information, and so is not unreasonable: in the model where the full Fisher information matrix exists, the nuisance parameter correction factor is $I_{12} I_{22}^{-1} I_{21}$. (See Section 2.1.) Thus in cases where the Fisher's matrix is diagonal, as in Model A, there is no correction for the unknown character of the nuisance parameter.

The analogy with the diagonal Fisher's information matrix is particularly appropriate because in many models the J-information is the corrected Fisher's information. In Type I Models A, B, and C the J-information is the same as the corrected Fisher information. This is quite generally true for densities which meet regularity conditions in θ and ϕ. The following theorem (4.2.5) has stronger assumptions than needed, but suffices to demonstrate the close relationship between J and I. But first, a lemma is needed.

4.2.3 Lemma. Suppose $\alpha \varepsilon R^u$ and $\beta \varepsilon R^v$ are row vectors. Suppose that I is an u+v by u+v symmetric positive definite matrix. Let I^{11} be the square submatrix in the upper left corner of I^{-1} of dimension u by u. Then

$$\inf_{\beta} (\alpha,\beta) I (\alpha,\beta)^t = \alpha (I^{11})^{-1} \alpha^t$$

and the infinum is attained at $\beta = \alpha I_{12} I_{22}^{-1}$.

Proof. Let I_{11}, I_{12}, I_{21}, and I_{22} be the submatrices corresponding to the natural partition of I between the u-th and u+1-th rows and the u-th and the u+1-th columns. From standard matrix results $(I^{11})^{-1} = I_{11} - I_{12} I_{22}^{-1} I_{21}$, so if we set

$$I_0 = \begin{bmatrix} I_{12} I_{22}^{-1} I_{21} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

then the quadratic form of interest becomes:

$$(11) \qquad (\alpha,\beta)\ I\ (\alpha,\beta)^t = (\alpha,\beta)\ I_0 (\alpha,\beta)^t + \alpha (I^{11})^{-1} \alpha^t$$

It is first shown that I_0 is nonnegative definite. Since I_{22} is real symmetric, there exists an orthogonal matrix P and a diagonal matrix D such that $I_{22} = PDP^t$ (see Rao {1973}, p. 39). From this it follows that $I_{22}^{-1} = PD^{-1}P^t$. For any row vectors $\alpha \varepsilon R^u$ and $\beta \varepsilon R^v$, define $\gamma := \alpha I_{12} PD^{-\frac{1}{2}}$ and $\delta := \beta PD^{\frac{1}{2}}$. Then

$$\begin{aligned} (\alpha,\beta)\ I_0 (\alpha,\beta)^t &= I_{12} I_{22}^{-1} I_{21} \alpha^t + 2\alpha I_{12} \beta^t + \beta I_{22}\ \beta^t \\ &= \gamma\gamma^t + 2\gamma\delta^t + \delta\delta^t \\ &= (\gamma+\delta)\ (\gamma+\delta)^t \end{aligned}$$

$$\geq 0$$

Now notice that if $\beta = -\alpha I_{12} I_{22}^{-1}$, then $-\alpha = \delta$ and so the first term on the right hand side of (11) is zero. Since it is always nonnegative, the infinum over β must be there attained, at which point (11) equals $\alpha(I^{11})^{-1}\alpha^t$.

For the parameter vector (θ, ϕ), let the superscript $^{(i)}$ denote partial differentiation with respect to the i-th component. Let $^{(i,j)}$ denote the second partial with respect to the i-th and j-th component.

4.2.4 Assumptions.

(a) Θ is a subset of R and Φ is a subset of R^s, for some integer s.

(b) The density $f(x; \theta, \phi)$ has for all x first and second partial derivatives in θ and ϕ at the null.

(c) $E(L^{(i)}(\theta_0, \phi_0)) = 0$ and $E(L^{(i,j)}(\theta_0, \phi_0)) = 0$.

(d) First and second order partial differentiation can be passed inside the expectation sign in $E L^2(\theta, \phi)$ at the null.

(e) There exists a solution $\xi(\theta)$ to

$$\inf_{\phi} \text{Var } L(\theta, \phi) = \text{Var } L(\theta, \xi(\theta))$$

in a neighborhood of θ_0 such that $\xi(\theta_0) = \phi_0$ and ξ is twice differentiable in θ at θ_0.

4.2.5 Theorem. Under assumptions 4.2.4, $J(\theta;\phi) = I(\theta;\phi)$.

Proof. First use 4.2.4.d to compute the partials of Var $(L(\theta,\phi))$ at the null:

$$(12) \qquad (\text{Var } L(\theta,\phi))^{(i)} = (E\ L^2)^{(i)} = 2\ E\ (L^{(i)}L)$$
$$= 0 \text{ at the null.}$$

And

$$(13) \qquad (\text{Var } L(\theta,\phi))^{(i,j)} = 2\ E\ L^{(i,j)}\ L + 2\ E\ L^{(i)}L^{(j)}$$
$$= 2\ I_{ij} \text{ at the null.}$$

We wish to compute the limit of $J^*(\theta)/(\theta-\theta_0)^2$ as θ approaches θ_0. Since $J^*(\theta) = \text{Var } (L(\theta,\xi(\theta))$, the above conditions ensure that $J^*(\theta)$ is twice differentiable at the null with respect to θ. An application of L'Hopital's rule may therefore be made, as in equation (7). Let $\zeta(\theta) = (\theta,\xi(\theta))$. Then

$$J^{*\prime}(\theta_0) = \sum_i \{\text{Var } L(\theta_0,\xi(\theta_0))\}^{(i)}\ \zeta_i'(\theta_0)$$
$$= 0 \text{ by (12) above.}$$

The second derivative is

$$J^{*''}(\theta_0) = \sum_i \sum_j (\text{Var } L(\theta_0, \xi(\theta_0))^{(i,j)} \zeta_i' \zeta_j' + \sum_i \text{Var } L(\theta_0, \xi(\theta_0))^{(i)} \zeta_i''$$

$$= \sum_i \sum_j 2\, I_{ij}\, \zeta_i' \zeta_j' \quad \text{by (12) and (13).}$$

It follows that with $\zeta' = (1, \xi'(\theta_0))$,

$$J^{*''}(\theta_0) = 2(\zeta')^t\, I\, (\zeta').$$

Now apply lemma 4.2.3, with $\alpha=1$ and $\beta=\xi'(\theta_0)$. The result is that $J^{*''}(\theta_0)$ would be minimized if $\xi'(\theta_0)=-I_{12}I_{22}^{-1}$ and there take on value $2(I^{11})^{-1}$. Hence $J(\theta_0;\phi_0)\geq(I^{11})^{-1}$ by equation (7). Notice that if we set $\gamma(\theta) = -I_{22}^{-1}I_{21}(\theta-\theta_0)$, then

(14) $$\text{Var } L(\theta, \gamma(\theta)) \geq \text{Var } L(\theta, \xi(\theta)) = J^*(\theta).$$

Dividing through (14) by $(\theta-\theta_0)^2$, and letting $\theta \to \theta_0$ the limiting equation is

$$(I^{11})^{-1} \geq J(\theta_0;\phi_0).$$

Hence $J(\theta_0;\phi_0) = (I^{11})^{-1} = I(\theta_0;\phi_0)$.

Next, it is shown that the J-information has the

computationally desirable property of additivity.

4.2.6 Lemma. (Additivity) Suppose X_1 and X_2 are independent observations from densities $f_i(x_i;\theta,\phi_i)$, $i = 1,2$. Suppose X_1 and X_2 have exact upper (lower) J-informations $J_1^+(\theta;\phi_1)$ and $J_2^+(\theta;\phi_2)$. Then the joint variables (X_1,X_2) have exact upper (lower) information $J^+(\theta;\phi_1,\phi_2) = J_1^+(\theta;\phi_1) + J_2^+(\theta;\phi_2)$.

Proof. Let $L_i = f_i(X_i;\theta,\phi_i)/f_i(X_i;\theta_0,\phi_{i0})$, $i = 1,2$, where $(\theta_0,\phi_{10},\phi_{20})$ is the null parameter at which the information is being computed. Then the likelihood ratio of (X_1,X_2) is L_1L_2, and

$$\text{(15)} \qquad \operatorname{Var} L_1L_2 = EL_1^2L_2^2 - 1 = E(L_1^2-1) + E(L_2^2-1) + E(L_1^2-1)\,E(L_2^2-1)$$

Let $J_i^*(\theta) := \inf_{\theta} E\,(L_i^2 - 1)$, $i = 1,2$. By taking infinums over (ϕ_1,ϕ_2) and dividing by $(\theta-\theta_0)^2$ the last equation becomes

$$\text{(16)} \qquad (\theta-\theta_0)^{-2} \inf_{(\phi_1,\phi_2)} \operatorname{Var} L_1L_2 = \frac{J_1^*(\theta)}{(\theta-\theta_0)^2} + \frac{J_2^*(\theta)}{(\theta-\theta_0)^2} + J_1^*(\theta)J_2^*(\theta)(\theta-\theta_0)^{-2}.$$

The exactness assumption implies that it is sufficient to consider limits as $\theta \to \theta_0^+$. If either one of J_1^+ or J_2^+ is infinite, then the limit as $\theta \to \theta_0^+$ of the right hand side of (16) is $+\infty$. If both the upper informations are finite, then

$$\lim_{\theta \to \theta_0^+} \quad J_1^*(\theta) J_2^*(\theta) \; / \; (\theta - \theta_0)^2 \quad = 0 \; J_1^+ \; J_2^+ = 0.$$

Hence from (16), in either case,

$$J^+(\theta_0; \phi_{10}, \phi_{20}) = J_1^+ + J_2^+ \; ,$$

as required.

4.2.7 Lemma. (Additivity) Suppose $X_1, \ldots, X_k$ are i.i.d. observations from density $f(x; \theta, \phi)$. If X has upper (lower) information $J_1^+(\theta; \phi)$ and $J_1^*(\theta) \to 0$ as $\theta \to \theta_0^{+(-)}$, then the joint upper (lower) information $J^+(\theta; \phi)$ in $X_1, \ldots, X_k$ is $kJ_1^+(\theta; \phi)$. (Exactness not required.)

Proof. Let $L_i := f(X_i; \theta, \phi) / f(X_i; \theta_0, \phi_0)$. Then $L := \Pi L_i$ is the joint likelihood ratio. Then $EL^2 = (EL_1^2)^k$, so

$$\inf_{\phi} \{EL^2\} = J^*(\theta) + 1 = \left(\inf_{\phi} \{EL_1^2\} \right)^k$$

$$= \left(J_1^*(\theta) + 1 \right)^k .$$

It follows that

$$\liminf \; J^*(\theta)/(\theta-\theta_0)^2$$

$$= \liminf \left((J_1^*(\theta)+1)^k - 1 \right) / (\theta-\theta_2)^2$$

$$= \liminf \frac{(J_1^*(\theta)+1)^k - 1}{J_1^*(\theta)} \quad \frac{J^*(\theta)}{(\theta-\theta_0)^2}$$

Since $J^*(\theta) \to 0$ as $\theta \to \theta_0^{+(-)}$, this then becomes,

$$= k \liminf J_1^*(\theta)/(\theta-\theta_0)^2$$

$$= k \, J_1^+ .$$

4.2.8 Example. (Model D) Combining 4.2.6 and 4.2.7 with equation (8) gives the information in Model D as:

(17) $$J(\theta;\phi) = nk/\theta^2$$

4.2.9 Example. (Model E) Combining the information in the pair (X_i, Y_i) by using 4.2.7, then using 4.2.6 on the set of n pairs, gives, with equation (10):

(18) $$J(\theta;\phi) = 2n/(\tfrac{1}{4}-\theta^2)$$

4.2.10 Comments. The J-information has been shown herein to be very similar to the Fisher's information. The advantage to this formulation is that for densities which are not differentiable in ϕ, one has an information measure very much like the Fisher's measure. It is additive, and, provided the number of nuisance parameters remains finite, it should be the basis of a tight lower bound for asymptotic variance. The J-information will in fact be the information of choice in those asymptotic models (such as Type II) in which the density for each observation (or likelihood factor) depends on the entire nuisance parameter. The discussion of J-information in the Type II model will come later in this chapter.

In any model, such as the Type I, in which each observation depends on an independent parameter, the J-information is, however, insufficient. Basing a lower bound on the J-information in Type I Model A, for example, leads to the same lower bound as Andersen's. This maintains the conflict between the bound and the intuitive notion that s^2_{nk} has all the information. By using the Barankin lower bound (introduced next), a better measure (K-information) will be derived. The K-information will be found to coincide with the J-information in the Type II model,

but will be smaller in the Type I model, where the J and I informations appear to be too large.

4.3 Barankin's Lower Bound. In a 1949 paper, E.W. Barankin found the best possible found for unbiased estimation for a large class of problems. This is the setting: The parameter space is Ω indexed by λ, with no given structure. Once again, there are densities with respect to a sigma finite measure ν. Assume that the likelihood ratio $L(\lambda) := f(x;\lambda)/f(x;\lambda_0)$ is defined almost everywhere (ν) and has finite second moment with respect to the null measure λ_0. The goal is to estimate a (non-constant) function $g(\lambda)$.

4.3.1 Definition. Define Barankin's lower bound for unbiased estimators of $g(\lambda)$ at λ_0 to be

$$C(g,\lambda_0) := \sup \{ \frac{\left| \sum_i a_i (g(\lambda_i) - g(\lambda_0)) \right|^2}{\mathrm{Var}(\sum_i a_i L(\lambda_i))} \},$$

where the supremum is taken over finite sets of real numbers $\{a_i\}$ and parameter points $\{\lambda_i\}$ such that the denominator and the numerator are not both zero.

4.3.2 Theorem. (Barankin {1949})

(i) A necessary and sufficient condition that there

exist an unbiased estimate of g with finite variance at is that $C(g,\lambda_0)<\infty$.

(ii) If T is unbiased for $g(\lambda)$, for all λ, then Var $T \geq C(g,\lambda)$, for all λ.

(iii) If $C(g,\lambda_0)<\infty$, then there exists a unique unbiased estimate T_0 with Var $T_0 = C(g,\lambda_0)$. Thus T_0 is the unique unbiased estimate which is best at λ_0. (Locally best unbiased)

4.3.3 Kiefer's Modifications. In 1951 Kiefer pointed out that the constant $C(g,\lambda)$ could be derived as follows. Suppose Ω is a measurable space, with all one point sets $\{\lambda\}$ included in the sigma-field of sets. Denote by $f(x;P)$ the mixed probability density $\int f(x;u)\, dP(u)$. Let $g(P) := \int g(u)\, dP(u)$ and $L(P) := \int L(u)\, dP(u)$. Then we have

$$(19) \qquad C(g) = \sup_{P_1,P_2} \{ (g(P_1) - g(P_2))^2 / E(L(P_1)-L(P_2))^2 \},$$

where the supremum is taken over probability measures P_1 and P_2 which have mass concentrated on a finite number of points. This reault is easily obtained from definition 4.3.1.

The power of formulation (19) comes from extending the class of probability measures as follows:

4.3.4 Definition. Define a probability measure P on Ω to be unbiased for g if there is a support Ω_0 for P such that $\lambda << \lambda_0$ for all $\lambda \in \Omega_0$ and if for all T unbiased for g,

$$E(T;P) = g(P). \tag{20}$$

It should be noted that (20) is equivalent to the following interchange of orders of integration:

$$\int\int T\, f(x;u)\, dP(u)\, d\nu(x) = \int\int T\, f(x;u)\, d\nu(x)\, dP(u). \tag{21}$$

This interchange can be justified under two conditions using Fubini's Theorem (see Loève {1977}, p. 137): If T is nonnegative; or if Tf is absolutely integrable $P \times \nu$. Use of Fubini's Theorem is justified only if Ω and the sample space are sigma-finite.

If P_1 and P_2 are measures unbiased for g, then application of the Cauchy-Schwartz inequality yields

$$E(T-g(\lambda_0))^2\, E(L(P_1) - L(P_2))^2 \geq (g(P_1)-g(P_2))^2. \tag{22}$$

Equation (22) yields the lower bound result

$$\text{Var } T \geq \sup_{P_1,P_2} (g(P_1)-g(P_2))^2/E(L(P_1)-L(P_2))^2, \tag{23}$$

where the supremum is taken over P_1, P_2 which are unbiased for g. It should be noted that Kiefer (1951) derived this bound without making the unbiased requirement (21) explicit. The advantages of Kiefer's bound (23) over Barankin's are twofold: First, it provides a bound even when the likelihood ratios L(.) are not defined almost everywhere-ν. Secondly, since the class of probability measures has been extended, it is often easier to find a measure at which the bound is attained. Both these points are reinforced by the following examples from Kiefer (1951).

4.3.5 Example. Let $X_1,\ldots,X_n$ be i.i.d. observations from a uniform $(0,\lambda)$ distribution, for $\lambda>0$. By sufficiency it suffices to consider the maximum Y of the observations. It can be shown that, as $n\to\infty$, the Chapman-Robbins bound becomes $.648\lambda_0^2/n^2$. On the other hand, if we define $dP_1(u) = (n+1)\lambda_0^{-1}(u/\lambda_0)^n\,du$ for $u\varepsilon(0,\lambda_0)$ and define P_2 by $P_2\{\lambda_0\}= 1$, then the Kiefer lower bound becomes $\lambda_0^2/n(n+2)$. It is attained by the unbiased estimator $Y(n+1)/n$.

4.3.6 Example. Let $X_1,\ldots,X_n$ be i.i.d. observations from the distribution with density $\exp(-(x-\lambda))$ for $x\geq\lambda$. Here Z, the minimum of the observations, is sufficient

for λ and has density $\exp(-n(z-\lambda))$ for $z \geq \lambda$. The Chapman-Robbins bound is $.648/n^2$. On the other hand, putting $dP_1(u) = n \exp(-n(u-\lambda_0))$ for $u \geq \lambda_0$ and letting P_2 put mass 1 at $\{\lambda_0\}$ the Kiefer lower bound is n^{-2}, attained by the variance of the unbiased estimator $Z = n^{-1}$.

These two examples suggest that significant improvements in lower bounds can be made by considering mixing distributions over the parameter space. The next step is to use this additional leverage to derive a new information measure.

4.4 K-information. We start by putting the Barankin lower bound in a form which resembles the Chapman-Robbins bound. Let $\mu\{\theta\}$ and $\mu\{\phi\}$ be probability measures with mass 1 at $\{\theta\}$ and $\{\phi\}$ respectively. Suppose that Q is a probability measure on Φ such that $\mu\{\theta\} \times Q$ is an unbiased measure on $\Theta \times \Phi$. Then using $P_1 = \mu\{\theta\} \times Q$ and $P_2 = \mu\{\theta\} \times \mu\{\phi_0\}$ in (22) yields

$$\text{(24)} \qquad \text{Var } T \text{ Var } L(\theta,Q) \geq (g(\theta) - g(\theta_0))^2,$$

where

$$L(\theta,Q) := f(X;\theta,Q)/f(X;\theta_0,\phi_0).$$

Comparing this with equation (1) suggests the following modification of the J* distance.

4.4.1 Definition. Let P_θ be the family of all probability measures Q on Φ such that $\mu\{\theta\} \times Q$ is unbiased. Define the K*-distance between θ and the null (θ_0,ϕ_0) to be

$$K^*(\theta) := \inf \{ \text{Var } L(\theta,Q) : Q \in P_\theta \}. \tag{25}$$

If Var $L(\theta,\phi)$ is a measure of the distance between (θ_0,ϕ_0) and (θ,ϕ), then K* is an indication of how much that distance can be reduced by choosing "least favorable distributions" on the nuisance parameter ϕ. The notion of least favorable distribution plays an important role in the theory of hypothesis testing, for which see Lehmann (1959), p. 90. In the hypothesis testing framework, one seeks the distribution Q which minimizes the power of the most powerful size α test of H : (θ,Q) versus K : (θ_0,ϕ_0). In (25), one seeks distributions which minimize the variance of the likelihood ratio.

For the purpose of computing K*, it will be desirable to extend the family P_θ to a larger family P_θ^*. The next lemma demonstrates that in many cases we may take the infinum over all probability measures on Φ.

4.4.2 Lemma. If $E(T;\theta,\phi)$ is continuous in ϕ for all positive measurable functions T with finite mean and if Φ_θ is the set of all ϕ's such that $(\theta,\phi)<<(\theta_0,\phi_0)$, then every probability measure with compact support in Φ_θ is in P_θ. Further, if P_θ^* is the family of tight probability measures with support in Φ_θ, then

$$K^*(\theta) = \inf \{ \text{Var } L(\theta,Q) : Q \in P_\theta^* \} .$$

Proof. See Billingsley (1968), p. 9, for a discussion of tightness. Herein we use the result that if Q is a tight measure, then there exists an increasing sequence of compact sets $\{A_k\}$ such that $Q(A_k)\to 1$ as $k\to\infty$.

Suppose T is an unbiased estimator of g. Suppose that Q has compact support $A\subset\Phi_\theta$. Then $E(|T|;\theta,\phi)$, by continuity, is a bounded continuous function of ϕ on that support. Because of the boundedness, Fubini's Theorem may be applied to make the following interchange of orders of integration:

$$g(\theta) = \int E(T;\theta,\phi)\, dQ = E(T;\theta,Q).$$

Hence $\mu\{\theta\} \times Q$ is unbiased.

If Q is tight, but not of compact support, choose compact set A such that $Q(A) \geq 1-\varepsilon$. Define the conditional

probability measure Q* by Q* (B) = Q(A∩B)/Q(A). We then have:

$$f(x;\theta,Q^*) = Q^{-1}(A)\ \{\ f(x;\theta,Q) - \int_{A^c} f(x;\theta,\phi)\ dQ\ \}$$

$$\leq f(x;\theta,Q)/Q(A) \tag{26}$$

Hence we have:

$$\begin{aligned} \text{Var}\ (\ L(\theta,Q^*)) &= E(\ f(X;\theta,Q^*)/f(X;\theta_0,\phi_0))^2 - 1 \\ &\leq E(L^2(\theta,Q)/Q^2(A)) - 1 \\ &\leq E(L(\theta,Q))^2\ (1-\varepsilon)^{-2} - 1. \end{aligned} \tag{27}$$

Now as ε becomes arbitrarily small, the right hand side of equation (27) approaches Var (L(θ,Q)). Hence

$$\lim_{A\uparrow\Phi_\theta} \text{Var}\ L(\theta,Q^*) \leq \text{Var}\ L(\theta,Q),$$

where $Q^* \varepsilon P_\theta$, and so $K^*(\theta) \leq \inf\{\text{Var}\ L(\theta,Q) : Q \varepsilon P_\theta^*\}$. The reverse inequality is trivial.

4.4.3 Definition. Define the upper K-information to be

$$K^+ := \lim_{\theta\to\theta_0^+} \inf\ K^*(\theta)/(\theta-\theta_0)^2\ ;$$

the lower K-information to be

$$K^- := \lim_{\theta \to \theta_0^-} \inf K^*(\theta)/(\theta-\theta_0)^2 \ ;$$

and the K-information to be

$$K := \inf \{ K^-, K^+ \}.$$

An upper or lower K-information will be called exact if "lim inf" may be replaced by "lim."

4.4.4 Theorem. (Lower Bound) If T is an unbiased estimator of g then

$$\text{Var } T \geq (g'(\theta))^2 / K(\theta;\phi)$$

Proof. Follows from (24 and (25).

4.4.5 Theorem. $K(\theta;\phi) \leq J(\theta;\phi)$.

Proof. The relationship $J^*(\theta) \geq K^*(\theta)$ follows by comparing (2) and (25).

Theorem 4.4.4 demonstrates that K-information serves the same role in the formation of lower bounds for unbiased

estimation as I and J-information. Theorem 4.4.5 indicates that it is possible that progress has been made in the direction of finding a more accurate measure. The possibility is confirmed in the following discussion of Model A. The K-information in other models will not be computed until necessary preliminary remarks concerning K-information and likelihood factorizations are made in Section 4.6.

4.4.6 Example. (Model A) The K-information is the sought-for measure by which the estimate s^2_{nk}, used in Model A, is fully efficient. First, the requirements of Lemma 4.4.2 are satisfied. The proof is postponed to 4.9.2. For $\theta \to \theta_0^-$, it will be shown that there is a sequence of probability measures Q_θ in P^*_θ such that

$$\lim_{\theta \to \theta_0^-} \text{Var } L(\theta, Q_\theta)/(\theta-\theta_0)^2 = n(k-1)/2\theta_0^2.$$

Then, because Q_θ is in P^*_θ,

$$K^-(\theta_0;\phi_0)^{(n)} \leq n(k-1)/2\theta_0^2. \tag{28}$$

On the other hand, Var $s^2_{nk} = \{n(k-1)/2\theta_0^2\}^{-1}$, so by Theorem 4.4.5

(29) $$K(\theta_0;\phi_0^{(n)}) \geq n(k-1)/2\theta_0^2.$$

Since $K \leq \bar{K}$, we have by (28) and (29) that

(30) $$K(\theta_0;\phi_0^{(n)}) = \bar{K}(\theta_0;\phi_0^{(n)}) = n(k-1)/2\theta_0^2.$$

It follows that the K-efficiency of s^2_{nk} is 1, provided Q_θ exists.

Now it simplifies the computations (and suggests a generalization) to work with the density f of the sufficient statistics $(S^2_{nk}, \bar{X}_{1\cdot}, \bar{X}_{2\cdot},\ldots,\bar{X}_{n\cdot})$. Because of independence this can be written

$$f(s,x^{(n)};\theta,\phi^{(n)}) = p(s;\theta)r(x_1;\theta,\phi_1)\ldots r(x_n;\theta,\phi_n),$$

where p, the marginal of S^2_{nk}, is the density of θ times a chi-square with n(k-1) degrees of freedom, and r_i, the marginal of $\bar{X}_{i\cdot}$, is the density of a normal random variable with mean ϕ_i and variance θ/k.

The probability measure to be used on the nuisance parameter $(\phi_1,\ldots,\phi_n)$ is defined by $Q_\theta = Q_{\theta 1}x\ldots xQ_{\theta n}$, where

(31) $$Q_{\theta i} \overset{L}{=} N(\phi_{i0},(\theta_0-\theta)/k).$$

Under (θ,ϕ_i), $\bar{X}_{i\cdot} \overset{L}{=} \phi_i + \eta(\theta)$, where $\eta(\theta)$ is normally

distributed with mean 0 and variance θ/k. Hence, if ϕ_i has the distribution in (31), then by the convolution formula for normal random variables, under $(\theta, Q_{\theta i})$

$$\bar{X}_{i.} \overset{L}{=} \phi_{i0} + \eta(\theta + (\theta_0 - \theta)) = \phi_{i0} + \eta(\theta_0).$$

That is, $\bar{X}_{i.}$ has the same distribution under $(\theta, Q_{\theta i})$ and (θ_0, ϕ_{i0}). In fact, if we use the usual smooth version of the normal density for likelihood r, then

$$(32) \qquad r(x_i; \theta, Q_{\theta i}) = \int r(x_i; \theta, \phi)\, dQ = r(x_i; \theta_0, \phi_{i0}).$$

Thus the likelihood ratio for the density f, which is

$$(33) \qquad L(\theta, Q_\theta) = \frac{p(S^2_{nk}; \theta)\, r(\bar{X}_{1.}; \theta, Q_{\theta 1}) \ldots r(\bar{X}_{n.}; \theta, Q_{\theta n})}{p(S^2_{nk}; \theta_0)\, r(\bar{X}_{1.}; \theta_0, \phi_0) \ldots r(\bar{X}_{n.}; \theta_0, \phi_{n0})},$$

by virtue of (32) becomes

$$(34) \qquad L_p(\theta) = p(S^2_{nk}; \theta) \,/\, p(S^2_{nk} : \theta_0),$$

the partial likelihood ratio for the density of S^2_{nk}. This is exactly the likelihood ratio for a one-parameter problem involving S^2_{nk} alone. Hence we have

$$\lim_{\theta \to \theta_0^-} \text{Var } (L(\theta, Q_\theta))/(\theta - \theta_0)^2 = I^p(\theta_0) = n(k-1)/2\theta_0^2,$$

where I^p is the Fisher's information in marginal of S^2_{nk}.

4.4.7 Comments. This example has several notable features. One is the way in which the marginal densities of the means were factored out of the problem. Section 4.6 below is devoted to discovering the relationships between likelihood factorizations and the K-information.

The lower bound result of Theorem 4.4.4 also establishes some new information about unbiased estimation of functions of θ. While the Cramer-Rao efficiency of s^2_{nk} is $(k-1)/k$, there could conceivably exist a function $g(\theta)$ other than θ and an unbiased estimator T of g such that

$$\text{eff}_{CR}T = \frac{(g'(\theta_0))^2 / (nk/2\theta_0^2)}{\text{Var } T} > (k-1)/k.$$

It is now seen that consideration of the K-information does not allow such a possibility, since it would make the K-efficiency of T greater than 1 and thereby violate Theorem 4.4.4.

Another key point of this example is that the mixing distributions Q_θ used very definitely required $\theta < \theta_0$, as can be seen in (31). The asymmetry of the K-information

in the sense of $K^+ \neq K^-$ will continue to be a striking feature through the rest of this paper.

Another important observation to be made concerns the additivity of the measure. From equation (30), it follows that the K-information about θ in k i.i.d. $N(\phi,\theta)$ random variables is $(k-1)/2\theta^2$. Thus the K-information is not additive over i.i.d. observations, which makes its computation more difficult.

On the other hand, it would appear from (30) that the K-information might be additive over the n distinct nuisance parameters. This possibility is studied in Section 4.5. But first, the relationship between J,K, and the Type II model is discussed.

4.4.8 Type II model. Suppose X is an observation from density $f(x;\theta,Q)$, where Q is an unknown probability distribution on parameter space Φ. It will be itself regarded as a parameter from the space P of probability measures on Φ. Let (θ_0,Q_0) be the null. In this case the J*-distance is

$$J^*(\theta) = \inf \{ \text{ Var } L(\theta,Q) \ ; \ Q \varepsilon P \ \}.$$

This should be compared with (25), the definition of K*. The difference between a Type I model K*-information and a

Type II model J*-information is in the null distributions allowed and the restriction in K* to unbiased probability measures.

In order to determine a K*-distance for the Type II model, mixing distributions λ are put on the nuisance parameters, the Q-measures. Notice, however, that this does not increase the original family of likelihoods:

$$L(\theta,\lambda) = \int \int L(\theta,\phi)\, dQ(\phi)\, d\lambda(Q)$$
$$= \int L(\theta,\phi)\, dQ^*(\phi) = L(\theta,Q^*)$$

where the measure Q* is defined by

$$Q^*(B) = \int\int_B dQ(\phi)\, d\lambda(Q) = \int Q(B)\, d\lambda.$$

Since the minimization of Var $L(\theta,Q)$ takes place over the same family of likelihoods,

$$\inf_{\lambda} \text{Var } L(\theta,\lambda) = K^*(\theta) = J^*(\theta) = \inf_{Q} \text{Var } L(\theta,Q) \tag{35}$$

The K and J-informations for more than one observation from the Type II model are not necessarily identical. A discussion of this issue (4.5.9) must await a discussion of the additivity of the K-information.

4.5 Additivity of the K-information. Let $X_1,\ldots X_n$ be a sequence of independent random variables, where the density of X_i is $f(x;\theta,\phi_i)$. Let the null parameter be $(\theta_0,\phi_0^{(n)}):=(\theta_0,\phi_{10},\ldots,\phi_{n0})$. In the following, numeric subscripts indicate restriction to the density of X_i, whereas the superscript (n) indicates use of the product density. For example, $P_{\theta i}$ represents the family of measures which are unbiased for the likelihood $f(x;\theta,\phi_i)$ at (θ_0,ϕ_{i0}), where $P_\theta^{(n)}$ represents the measures on Φ^n which are unbiased for the joint density.

4.5.1 Lemma. Suppose that $P_{\theta 1}x\ldots xP_{\theta n}$ is contained in $P_\theta^{(n)}$. Then if X_i has exact upper (lower) K-information $K_i^+:=K^+(\theta_0;\phi_{i0})$, then the upper (lower) K-information in $(X_1,\ldots,X_n)$ satisfies

$$K^{+(n)}(\theta_0;\phi_0^{(n)}) \leq K_1^+ + \ldots + K_n^+. \tag{36}$$

Proof. It suffices to prove the lemma for n = 2. If L_i, i=1,2, is the likelihood ratio for X_i, then the joint likelihood ratio is $L = L_1L_2$. For any product measure Q_1xQ_2 on Φ x Φ ,

$$\begin{aligned}\text{Var } L(\theta,Q_1xQ_2) &= E(L^2(\theta,Q_1xQ_2) - 1)\\ &= E\ L_1^2(\theta,Q_1)E\ L_2^2(\theta,Q_2) - 1\end{aligned}$$

$$= \text{Var } L_1^2(\theta,Q_1) + \text{Var } L_2(\theta,Q_2)$$

$$+ \text{Var } L_1(\theta,Q_1)\text{Var } L_2(\theta,Q_2).$$

Taking the infimum over the measures Q_1 and Q_2 in the above gives

$$(37) \qquad \inf_{Q_1,Q_2} \text{Var } L(\theta,Q_1 x Q_2) = K_1^*(\theta) + K_2^*(\theta) + K_1^*(\theta)K_2^*(\theta)$$

However, the K*-distance for the joint likelihood is the infinum of the left hand side of (37) over all probability measures in $P_\theta^{(2)}$, not just the product measures. Hence

$$(38) \qquad K^{*(2)}(\theta) \leq K_1^*(\theta) + K_2^*(\theta) + K_1^*(\theta)K_2^*(\theta)$$

Divide equation (38) through by $(\theta-\theta_0)^2$ let $\theta \to \theta_0^+$, and use the exactness (the limits exist) of the upper K-informations to attain the result.

Lemma 4.5.1 establishes a form of subadditivity for the K-information. The full additivity of the information is true under some assumptions that result from a consideration of the minimization process in (25). The following definition will be useful for the discussion:

4.5.2 Definition. A family $\mathcal{P}$ of probability measures will be called convex if for all P and Q in $\mathcal{P}$, and any q in (0,1), the probability measure qP + (1-q)Q is in $\mathcal{P}$.

4.5.3 Lemma. $\mathcal{P}_\theta$ is convex.

Proof. $$\begin{aligned} E(T;\theta,qP+(1-q)Q) &= qE(T;\theta,P)+(1-q)(T;\theta,Q) \\ &= qg(\theta) + (1-q)\, g(\theta) \\ &= g(\theta). \end{aligned}$$

4.5.4 Definition. If there exists a probability measure P in $\mathcal{P}_\theta$ such that $\inf \{ EL^2(\theta,Q) ; Q\varepsilon\mathcal{P}_\theta \} = EL^2(\theta,P)$, then call P a least favorable distribution (LFD). Denote the set of all such least favorable distributions $LFD(\mathcal{P}_\theta)$.

4.5.5 Theorem.

(a) P is in $LFD(\mathcal{P}_\theta)$ if and only if $EL^2(\theta,P) \leq EL(\theta,P)L(\theta,Q)$ for all Q in $\mathcal{P}_\theta$.

(b) If P and Q are in $LFD(\mathcal{P}_\theta)$, then $L(\theta,Q) = L(\theta,P)$ a.s.-(θ_0,ϕ_0).

Proof. For any P and Q in $\mathcal{P}_\theta$ let F_q be the probability measure in $\mathcal{P}_\theta$ defined by $F_q := qP + (1-q)Q$. Let

(39) $$L_q := L(\theta,F_q) = qL(\theta,P) + (1-q)L(\theta,Q),$$

By the convexity of the square function

$$EL_q^2 \leq q\ EL^2(\theta,P) + (1-q)\ EL^2(\theta,Q), \tag{40}$$

with equality if and only if $L(\theta,P) = L(\theta,Q)$ a.s. (θ_0,ϕ_0). The proof of (b) is immediate from this.

As a function of q, EL_q^2 is a convex quadratic, so it has a unique minimum for $q\varepsilon(-\infty,\infty)$. There are three possible locations for the minimum:

(i) If EL_q^2 is minimized for $q \leq 0$, then it is strictly increasing for q in (0,1,) so $\inf\{EL_q^2; q\varepsilon(0,1)\} = EL^2(\theta,P)$. The function EL_q^2 is minimized for $q \leq 0$ if and only if

$$EL^2(\theta,P) \leq EL(\theta,P)\ EL(\theta,Q). \tag{41}$$

(ii) If EL_q^2 is minimized for $q \geq 1$, then on the range (0,1) the infinum is $EL^2(\theta,Q)$ and equation (41) holds with P and Q reversed in roles.

(iii) If EL_q^2 is minimized for q in (0,1), then the infinum occurs at

$$q = E\{L(\theta,P)L(\theta,Q) - L^2(\theta,P)\}/E\{L(\theta,P)-L(\theta,Q)\}^2 \tag{42}$$

and equals

$$(43) \qquad \frac{EL^2(\theta,P)\ EL^2(\theta,Q) - \{EL(\theta,P)L(\theta,Q)\}^2}{E\ \{L(\theta,P) - L(\theta,Q)\}^2}$$

Part (a) of the theorem is now proved by noting that if P is at least favorable distribution, then EL_q^2 must be minimized at $q \leq 0$ for all p.m. Q, and so (41) holds.

4.5.6 Theorem. (Existence) Suppose Φ is a separable metric space and let a be a compact subset of Φ. Let $P_\theta(a)$ be the subset of P_θ consisting of measures with support in a. Assume Lemma 4.4.2 holds. Suppose $EL(\theta,\phi_1)L(\theta,\phi_2)$ is finite and continuous in ϕ_1 and ϕ_2. Then there exists a least favorable distribution for $P_\theta(a)$.

Proof. Let $P_1, P_2, \ldots$ be a sequence of p.m. from $P_\theta(a)$ such that

$$(44) \qquad EL^2(\theta,P_n) \to \inf\{\ EL^2(\theta,Q) : Q\varepsilon P_\theta(a)\ \} \text{ as } n\to\infty.$$

The family $\{P_n\}$ is tight, as all measures have the same compact support, and it is therefore relatively compact (see Billingsley {1968}, p. 37). It follows that there exists a subsequence, labelled $\{P_n\}$ without loss of generality, and a probability measure P_0 such that $P_n \xrightarrow{w} P_0$ as $n\to\infty$.

By lemma 4.4.2, P_0 is in $P_\theta(a)$. This weak convergence,

along with separability, implies the weak convergence of the product measures, $P_n \times P_n \xrightarrow{W} P_0 \times P_0$ as $n \to \infty$. Now by assumption $EL(\theta, \phi_1)L(\theta, \phi_2)$ is continuous on $a \times a$, and hence bounded, since $a \times a$ is compact, so weak convergence implies

$$(45) \qquad EL^2(\theta, P_n) = \int\int EL(\theta, \phi_1)L(\theta, \phi_2)\, dP_n \times P_n$$

$$\to \int\int EL(\theta, \phi_1)L(\theta, \phi_2)\, dP_0 \times P_0 \text{ as } n \to \infty$$

$$= EL^2(\theta, P_0)$$

Comparing (44) and (45) gives the result.

4.5.7 Corollary. Under the same assumptions as in Theorem 4.5.6, if a_n is an increasing sequence of compact sets such that $a_n \to \Phi$ and if $P_n = LFD\{P_\theta(a_n)\}$, then

$$\lim_{n \to \infty} EL^2(\theta, P_n) = K^*(\theta) + 1.$$

Proof. For any Q in P_θ, let Q_n be the conditional probability measure $Q(. \mid a_n)$. Now $Q(a_n) \to Q(\Phi) = 1$, so, using equation (26), we have

$$(46) \qquad EL(\theta, P_n) \leq EL^2(\theta, Q_n) \leq EL^2(\theta, Q)/Q(a_n) \to EL^2(\theta, Q)$$

as $n \to \infty$. Since $K^*(\theta)+1$ is the infinum of the right hand

side of (46), we are done.

4.5.8 Theorem. (Additivity) In addition to the assumptions of Theorem 4.5.6, assume Φ is sigma-compact. Suppose that for each X_i in $X_1, X_2, \ldots, X_n$, a sequence of independent random variables, there is an exact upper K-information $K_i^+ := K(\theta_0; \phi_{i0})$ as in 4.5.1. Then the exact upper K-information in $(X_1, \ldots, X_n)$ is

$$K^{+(n)} = K_1^+ + \ldots + K_n^+.$$

Proof. It suffices to prove the theorem for n = 2. Let $L_i(\phi_i) := L(X_i; \theta, \phi_i)$ be the likelihood ratio for X_i, i = 1,2. Let a_k be an increasing sequence of compact sets such that $a_k \to \Phi$ as $k \to \infty$. Let P_{ki} be a least favorable distribution for likelihood L_i in family $P_{\theta i}(a_k)$. Then it will be shown that $P_{k1} \times P_{k2}$ is a least favorable distribution for likelihood $L = L_1 L_2$ in family $P_\theta^{(2)}(a_k \times a_k)$ by using the criterion of Theorem 4.5.5(a). By that theorem,

$$(47) \qquad EL_i^2(\theta, P_{ki}) \leq EL_i(\theta, P_{ki})\, L_i(\theta, \phi_i) \quad \text{for } i = 1,2,$$

where ϕ_i can take on any value in a_k. Equation (47) then yields

$$(48) \qquad E\, L_1^2(P_{k1})EL_2^2(P_{k2}) \leq \int \prod_{i=1}^{2} EL_i(P_{ki})L_i(\phi_i)dQ(\phi_1,\phi_2)$$

for any probability measure Q on $a_k \times a_k$. Rearranging terms on the right side of (48) and relabelling the expression on the left gives

$$(49) \qquad E\, L^2(P_{k_1} \times P_{k_2}) \leq E\, L(P_{k_1} \times P_{k_2})\, L(Q).$$

By Theorem 4.5.5(a), $P_{k_1} \times P_{k_2}$ is least favorable. From the corollary, 4.5.6,

$$\begin{aligned} K^{*(2)}(\theta) &= \lim_{k\to\infty} E\, L_1^2(P_{k_1})\, L_2^2(P_{k_2}) - 1 \\ &= \lim_{k\to\infty} E\, L_1^2(P_{k_1})\, E\, L_2^2(P_{k_2}) - 1. \end{aligned}$$

Now applying the corollary again to the right hand side, we have

$$(50) \qquad \begin{aligned} K^{*(2)}(\theta) &= (K_1^* + 1)(K_2^* + 1) - 1 \\ &= K_1^* + K_2^* + K_1^* K_2^*. \end{aligned}$$

Dividing (50) by $(\theta-\theta_0)^2$ and letting $\theta \to \theta_0^+$, the result is obtained.

The investigation of the additivity of the K-infor-

mation now provides the background needed for continuing the discussion started in Section 4.4.8 concerning the relationship between J and K-information in the Type II model.

4.5.9 Type II model. In the Type II model the K* and J* distances for a single observation X_i were identical (35). If, however, there are n i.i.d. observations $X_1,\ldots,X_n$ with null parameter (θ_0,Q_0), then

$$J^{*(n)}(\theta) = \inf_Q EL_1^2(Q)\ldots EL_n^2(Q) - 1$$

$$= \inf_Q \int EL_1^2(\phi_1)\ldots EL_n^2(\phi_n)\, dQxQx\ldots xQ-1 \tag{51}$$

whereas the K*-distance is

$$K^{*(n)}(\theta) = \inf_\lambda E\{\int L_1(Q)\ldots L_n(Q)\, d\lambda(Q)\}^2 - 1$$

$$= \inf_{Q^*} E\{\int L_1(\phi_1)\ldots L_n(\phi_n)\, dQ^*(\phi^{(n)})\}^2-1, \tag{52}$$

where the probability measure Q* is defined on ϕ^n by

$$dQ^*(\phi^{(n)}) := dQ(\phi_1)\ldots dQ(\phi_n)\, d\lambda(Q).$$

In comparing (51) and (52) one notices that the K* distance involves minimizing over measures Q* which may not

be true product measures. However, one of the important principles underlying the proof of Theorem 4.5.8 (additivity) is that in order to minimize $E(\int L_1(\phi_1)\ldots L_n(\phi_n)dQ)^2$, it typically suffices to use product measures $Q_1 x \ldots x Q_n$ where Q_i "minimizes" $EL_i^2(Q_i)$. An informal application of this principle to (51) and (52) yields the conclusion that for the Type II model the distance measures J* and K* are often identical. From this point forth only the J* measure will be used on Type II models.

This section has dealt with the additivity of the K*-information. Although these results are of mathematical interest, as far as estimation is concerned we are not merely interested in the K-information per se, but also in determining whether or not an estimation procedure is in some sense fully efficient. The next two sections examine this issue in terms of likelihood factorization and the K-lower bound (4.4.4).

4.6 Likelihood factorizations. Recall that in Chapters 2 and 3 likelihood factorizations were used as a basis for inference in the presence of nuisance parameters. These factorization schemes, including the Cox partial likelihood of Section 2.9, will now be put in a more general framework. The motivation for the following formal definitions

of likelihood factorizations is that they seem to capture some relevant mathematical features of likelihood factorizations without being tied to any particular mode of derivation. Thus we may compare at one time all marginal, conditional, Cox partial, and any other likelihood factorizations for a single density.

4.6.1 Definition. Suppose that the vector-valued random variable X has density $f(x;\theta,\phi)$ with respect to a sigma-finite measure. A function $g(x;\theta,\phi)$ will be called a likelihood if for all (θ,ϕ) and (θ_0,ϕ_0) in $\Theta\times\Phi$

$$E\{ L_g(\theta,\phi) ; \theta_0,\phi_0 \} = 1$$

where $L_g = g(X;\theta,\phi)/g(X;\theta_0,\phi_0)$ is the g-likelihood ratio.

4.6.2 The Cox likelihood. Suppose that for random vector $(Y_1,\ldots,Y_n)$ there exists for each i, $i = 1,\ldots,n$, and each fixed $y^{(i-1)}$, a measure $\nu_i(\cdot \mid y^{(i-1)})$ on the range space of Y_i such that the function $f_i(y_i \mid y^{(i-1)};\theta,\phi) = f_i(\theta,\phi)$ is the density with respect to ν_i for the conditional distribution of Y_i given $Y^{(i-1)} = y^{(i-1)}$ under (θ,ϕ). Then the joint density for the full vector $(Y_1,\ldots,Y_n)$ is $\prod_{i-1}^{n} f_i(\theta,\phi)$ with respect to the measure $\nu^{(n)}$ defined by

$$d\nu^{(n)}(y^{(n)}) = d\nu_n(y_n|y^{(n-1)})\ldots d\nu_1(y_1).$$

Under an obvious relabelling this can be seen to be a factorization of the Cox type (Section 2.9, Equation (8).) Let S be any nonempty subset of (1,2,...,n). Let $g = \Pi(f_i\ ;\ i\varepsilon S)$. It will be shown that g is a likelihood. Let (θ,ϕ) and (θ_0,ϕ_0) be given. Then

$$E(L_g\ ;\ \theta_0,\phi_0) = \int L_g \prod_i f_i(\theta_0,\phi_0)d\nu_n\ldots d\nu_1$$

$$= \int \prod_i f_i^*\, d\nu_n\ldots d\nu_1.$$

where $f_i^* := f_i(\theta_0,\phi_0)$ if $i \notin S$,

$:= f_i(\theta,\phi)$ if $i\varepsilon S$.

But for each i and each fixed $y^{(i-1)}$,

$$\int f_i^*\ (\ y_i\ |\ y^{(i-1)})\ d\nu_i(y_i|\ y^{(i-1)}) = 1$$

Hence

$$E\ (\ L_g\ ;\ \theta_0,\phi_0) = \int \prod_i\ f_i^*\, d\nu_n\ldots d\nu_1$$

$$= \int (f^*\ldots f_{n-1}^*)(\int f_n^* d\nu_n)d\nu_{n-1}\ldots d\nu_1$$

$$= \int (f^*\ldots f_{n-1}^*)(\int f_{n-1}^* d\nu_{n-1})d\nu_{n-2}\ldots d\nu_1$$

$$\vdots$$

$$= \int f_1^* \, d\nu_1 = 1,$$

as claimed.

The conclusion is that the Cox partial likelihood is indeed a likelihood in the sense of definition 4.6.1, where $S = \{2,4,6,\ldots\}$.

4.6.3 Definition. A factorization of the density $f = pr$ will be called a likelihood factorization if each of p and r are themselves likelihoods. It will be called a partial likelihood factorization if, in addition, one of p or r does not depend functionally on ϕ. It will be standard notation herein for this factor to be labelled p. It will be called the partial likelihood and r will be called the remainder likelihood.

4.7 Complete Factorizations. Given a likelihood factorization f=pr, how does one know if the "information" in r about θ is so confounded with nuisance parameters that it can be safely ignored for estimation purposes? The answer depends, of course, on the meaning given to the word "information." The K-information is used in this section to develop and support an answer to this question.

Recall from Chapter 3 the definitions of the partial score

$$U_p(X_i;\theta) = \partial \ln p(X;\theta)/\partial\theta,$$

the partial likelihood ratio,

$$L_p(X_i;\theta) = p(X;\theta)/p(X;\theta_0),$$

and the partial Fisher information

$$I_p(\theta;\phi) = E(\ U_p^2(\theta)\ ;\ \theta,\phi\ ).$$

This terminology will now be used for the definition of likelihood of 4.6.1. Notice, in particular, that if $|(L_p(\theta) - 1)/(\theta-\theta_0)|$ is dominated by integrable random variable Y in a θ-neighborhood of θ_0, then the likelihood definition of 4.6.1 implies that the mean of the score is zero:

$$(53) \qquad 0 = E(\ (L_p(\theta)-1)/(\theta-\theta_0)) \to E(\ U_p(\theta_0)\ ).$$

Hence $I_p = \text{Var } U_p$. If the dominating random variable Y is square integrable, then

(54) $$E\,(L_p(\theta) - 1)^2/(\theta-\theta_0)^2 \to E\,U_p^2 = I_p \text{ as } \theta\to\theta_0.$$

What is the relationship between the partial information and the K-information? The intuitive feeling is that the partial likelihood has less information than the full likelihood, hence

(55) $$I_p \leq K_f.$$

The following two lemmas reveal that (55) is generally true for factorizations where one of p or r is a marginal distribution. It is conjectured that it holds for all partial likelihood factorizations, subject to mild regularity conditions on p and r.

4.7.1 Lemma. If f = pr is a partial likelihood factorization such that p is a marginal distribution for some statistic C and p satisfies (54), then $K_f \geq I_p$.

Proof. Let $L_p := p(c;\theta)/p(c;\theta_0)$, $L_r := r(x|c;\theta,\phi)/r(x|c;\theta_0,\phi_0)$, and $L_r(Q) := \int L_r dQ(\phi)$. Then the variance of the likelihood ratio may be decomposed as follows:

$$E\,L_p^2L_r^2 - 1 = E\,L_p^2 - 1 + E\,L_p^2(L_r-1)^2 + 2EL_p^2(L_r-1).$$

It is shown that the last summand is zero. Write

$$E\ L_p^2(L_r-1) = E\ \{\ E\ (L_p^2(L_r-1)|C)\ \}$$

$$= E\ \{\ L_p^2\ E(L_r-1|C)\ \}$$

$$= E\ \{\ L_p^2 \times 0\ \} = 0.$$

In the above $E(L_r-1|C) = 0$ because L_r is the conditional likelihood ratio. Hence:

$$K^*(\theta) = E\ L_p^2 - 1 + \inf_Q E\ L_p^2\ (L_r(Q) - 1)^2.$$

The last summand is nonnegative, and equals zero if $L_r(Q)=1$. If (54) holds, dividing through by $(\theta-\theta_0)^2$ and letting $\theta\to\theta_0$ yields the desired result.

4.7.2 Lemma. If f = pr is a partial likelihood factorization such that (54) holds,

$$\lim_{\theta\to\theta_0} E(L_p-1)^4/(\theta-\theta_0)^4 = E(U_p)^4 < \infty$$

and r is a marginal likelihood for some statistic C, then $K_f \geq I_p$.

Proof. The same notation will be used as in Lemma

4.7.1. The variance of the likelihood ratio is decomposed

$$E\ L_p^2 L_r^2 - 1 = E\ L_p^2 - 1 + E\ L_r^2 - 1 + R,$$

where the remainder R may be expressed

$$\begin{aligned} R &= E(L_p^2-1)\ (L_r^2-1) \\ &= E(L_p-1)^2(L_p-1)^2 + 2E(L_p-1)^2(L_r-1) \\ &\qquad + 2E(L_p-1)(L_r-1)^2 + 4E(L_p-1)(L_r-1). \end{aligned}$$

The last two summands are zero. The last is zero because p and r are likelihoods. The next to last is zero because L_p is the conditional likelihood ratio (same argument as Lemma 4.7.1). Hence the variance of the likelihood ratio may be written

$$E\ L_p^2 L_r^2 - 1 = E\ L_p^2 - 1 + E(L_p-1)^2(L_r-1)^2 + R_1,$$

where $R_1 = E(L_r-1)^2 + 2\ E(L_p-1)^2(L_r-1)$. Let $R_1(H_\theta)$ be the corresponding variable with $L_r(H_\theta)$ substituted for L_r. It will be shown that

$$\liminf_{\theta \to \theta_0^{+}(-)} R_1(H_\theta)\ (\theta-\theta_0)^{-2} \geq 0$$

for any sequences of distributions H_θ. The conclusion of the lemma will follow because $E(L_p-1)^2(L_rH_\theta-1)^2 \geq 0$.

Because of the inequality $(a+b)^2 \geq 0$, for any real numbers a and b, the following inequality holds:

$$-2(L_p-1)^2(L_r-1) \leq (L_p-1)^4+(L_r-1)^2.$$

So for any H_θ,

$$\liminf_{\theta\to\theta_0^{+(-)}} R(H_\theta)(\theta-\theta_0)^{-2} \geq \lim_{\theta\to\theta_0} - E(L_p-1)^4(\theta-\theta_0)^{-2}$$

$$= 0$$

by assumption.

The two preceding lemmas are motivation for the notion of a complete factorization, where the word complete is meant to suggest that the information about θ has been completely factored into the partial likelihood.

4.7.3 Definition. Suppose that for the partial likelihood factorization f = pr it is true that

(56) $$K_f = I_p.$$

where K_f is the K-information in density f and I_p is the

partial Fisher's information in p. Then the factorization will be called complete.

4.7.4 Theorem. If f = pr is a complete factorization, then for any T unbiased for g

$$\text{Var } T \geq (g'(\theta_0))^2 / I_p. \tag{57}$$

Proof. Compare (55) and the K-lower bound 4.4.4.

This theorem suggests that if "all the information" in p can be used by a statistic, then the bound is tight. The simplest case occurs if p is a marginal distribution for a statistic C. Then (57) is the Cramer-Rao lower bound for unbiased estimates T which are functions of C alone.

4.7.5 Corollary. If f = pr is a complete factorization, with p a marginal likelihood, then any unbiased statistic which is a linear function of the partial score U_p will meet the lower bound.

4.7.6 Applications. In Model A, the partial is the marginal of S^2_{nk}, which is θ times a chi-square with n(k-1) degrees of freedom. Hence

$$U_p = \tfrac{1}{2}(-n(k-1)\theta^{-1}) + \tfrac{1}{2}\,\theta^{-2}\,S^2_{nk}.$$

Thus s^2_{nk} meets the lower bound.

In Model D, the exponentials with unknown support, the scaled spacings Z_{ij} are i.i.d. exponential, mean θ, random variables. (See 2.6.3.) The partial score, based on their marginal distribution, is

$$U_p = -n(k-1)\theta^{-1} + \theta^{-2}Z..,$$

so $\overline{Z}..$ meets the lower bound.

In Model E, the genetics Bernoulli, the marginal likelihood of the W_i is that of independent Bernoullis with common mean $q = (\tfrac{1}{2} - 2\theta^2)$. If the problem is transformed into the q-parameter system (one to one), then

$$U_p(q) = (W. - nq) \,/\, q(1-q).$$

Hence $\overline{W}.$, as an estimate of q, meets the lower bound.

4.7.7 Conditional likelihoods. The lower bound (57) does not have a simple interpretation for purely conditional partial likelihoods, such as in Models B and C, the paired binomials. The motivation for the definition 4.7.1 for conditional likelihoods must wait until Chapter 5,

when lower bounds for asymptotic variance are derived based on a modified version of K_f. This will imply that if an estimator can be found with asymptotic variance equal to the Cramer-Rao type lower bound based on I_p, as was true for the CMLE (3.6.4), then complete factorization will ensure that the bound is tight and that those estimators are optimal in that setting.

The following lemma demonstrates that any lower bounds for unbiased estimation derived for Models B and C must serve merely as guidelines to the intuition.

4.7.8 Lemma. In Models B and C there are no unbiased estimates for any nonconstant function of θ.

Proof. Let $C = X + Y$. Then for all θ, $P(C = 0;\theta,\phi) \to 1$ as $\phi \to -\infty$. Also, when $C = 0$, the conditional distribution of X given C does not depend on θ, so $E(\ T \mid C = 0)$ is a constant for any statistic T. Now notice that

$$E\ T = \sum_c E(T|C=c;\theta)\ P(C=c;\theta,\phi) \to E(T|C=0) \text{ as } \phi\to-\infty.$$

Hence T cannot be unbiased.

4.8 Simple Nuisance Likelihoods. If it were always a simple matter to compute the information K_f, then checking

for completeness of the factorization would be straightforward. However, it appears to be difficult to directly compute the infinum of Var $L(\theta,Q)$ over P_θ. For that reason, a sufficient condition is developed in this section.

In example (4.4.6), instead of directly computing $K^*(\theta)$ for Model A by finding the infinum of Var $L(\theta,Q)$, we used a sequence of mixing distributions Q_θ in Var $L(\theta,Q)$ which appeared to "least favorable" in the sense that they eliminated the remainder likelihood from the likelihood ratio, so that $L(\theta,Q_\theta) = L_p(\theta)$. This approach is now formalized so that it may be used on the other models.

4.8.1 Definition. Suppose that r is a likelihood such that for either $\theta\to\theta_0^+$ or $\theta\to\theta_0^-$, there exist probability measures Q_θ in P_θ^* such that

$$\int r(x;\theta,\phi)\, dQ_\theta(\phi) = r(x;\theta_0,\phi_0) \text{ for all } x. \tag{58}$$

Then r is called a simple nuisance likelihood. It will be called upper or lower depending on whether $\theta\to\theta_0^+$ or $\theta\to\theta_0^-$ in (58). Distributions which satisfy (58) will be called nuisance eliminating.

4.8.2 Lemma. Suppose that f = pr is a partial likelihood factorization such that the partial information I_p exists,

(54) holds, and r is a simple nuisance likelihood. If $K_f \geq I_p$, then pr is a complete factorization.

Proof. By the definition of K*-distance,

$$K^*(\theta) \leq E\,(L(\theta, Q_\theta) - 1)^2 = E\,(\,L_p(\theta) - 1\,)^2.$$

Divide through by $(\theta-\theta_0)^2$ and let $\theta \to \theta_0$.

4.8.3 Comments. The use of the word "simple" in simple nuisance likelihood is meant to suggest that showing that (58) holds is simple relative to the difficulty in computing K-information. As the examples in Section 4.9 will demonstrate, demonstrating that there exist Q_θ such that (58) holds is often nontrivial.

The word "simple" is appropriate in another sense. In a partial likelihood factorization pr, if the remainder r is a simple nuisance, then the factorization possesses the following uniqueness property.

4.8.4 Lemma. If f = pr is a partial likelihood factorization and r is a simple nuisance, then the factorization is unique in the sense that for any other such factorization p*r*, $L_p^{(\theta)} = L_{p^*}^{(\theta)}$ a.s.-(θ_0, ϕ_0) for all θ such that (58) holds for r and r*. Further, E ln L_p is maximized

among all partial likelihood factorizations by the factorization where r is a simple nuisance.

Proof. Suppose the p*r* is a partial likelihood factorization, but r* is not necessarily simple nuisance. Then

(59) $\quad pr = p^*r^* \text{ implies } r = p^*r^*/p.$

Since r is a simple nuisance, the equality $r(x;\theta,Q_\theta) = r(x;\theta_0,\phi_0)$ holds for some p.m. Q_θ. The conclusion of (59) implies

$$r((x;\theta,Q_\theta)^{p^*}(x;\theta)/p(x;\theta) = r^*(x;\theta_0,\phi_0)p^*(x;\theta_0)/p(x;\theta),$$

which becomes, with rearranging,

$$L_{r^*}(\theta,Q) = L_p(\theta)/L_{p^*}(\theta).$$

Since r* is a likelihood, $E\ L_{r^*}(\theta,Q) = 1$, and so application of Jensens inequality (e.g. Loeve {1977}, p. 161) yields

(60) $\quad 0 = \ln EL_{r^*} \leq E \ln L_{r^*} = E \ln L_p(\theta) - E \ln L_{p^*}(\theta),$

with equality if and only if $L_p(\theta) = L_{p*}(\theta)$ a.s.-null. Both conclusions of the lemma now follow.

Let us now consider how the simple nuisance criterion applies to the Type II model. Suppose that the density f for each of the i.i.d. random variables X_i has a partial likelihood factorization $f(x;\theta,Q) = p(x;\theta)\ r(x;\theta,Q)$. Recall from Section 4.4.8 that putting a probability measure λ on the space of probability measures on Φ did not increase the family of likelihoods. Thus for the Type II model the proper interpretation of (58) is that for null (θ_0,Q_0) and $\theta\to\theta_0^+$ or $\theta\to\theta_0^-$, there exist probability measures Q_θ on Φ such that

(61) $$r(x;\theta,Q_\theta) = r(x;\theta_0,Q_0) \text{ for all } x.$$

We are fortunate in that once (58) is shown for the original space Φ, (61) follows:

4.8.5 Lemma. Suppose that for each ϕ in Φ there exists a probability measure Q_θ^ϕ such that

$$r(x;\theta,Q_\theta^\phi) = r(x;\theta_0,\phi) \text{ for all } x.$$

Then for each probability measure Q_0 there exists a

probability measure Q such that (61) holds.

Proof. Let $dQ_\theta(y) = dQ_\theta^\phi(y)\, dQ_0(\phi)$.

The reverse implication does not follow. That is, it is possible that (61) holds for a large family of nulls Q_0 without (58) holding. This will be the case for Model C (Section 4.9). The significance of this observation is that we will be able to draw strong lower bound conclusuions only for the Type II model in the paired binomial problem.

To summarize Section 4.8, the central topic was the concept of the simple nuisance likelihood. It was shown to imply complete factorization, given regularity of the partial likelihood. A nuisance likelihood could thus be said to have no K-information. The uniqueness properties shown to be implied by the simple nuisance criterion provide reassurance that the nuisance criterion will not lead to dual routes of inference. Finally, the notion of nuisance likelihood carried over in a simple manner into the Type II model. The next step, and the goal of Section 4.9, is to show that the simple nuisance likelihood criterion can be applied to the key models to determine if a factorization is complete.

4.9 The Key Models Revisited. It is time (at long last) to return to the five models introduced in Chapter 2. Factorizations in each model were displayed there which seemed to partition out most of the information about θ. We now examine those factorizations in light of the results of this chapter and determine if the remainder likelihoods of the simple nuisance type.

Before commencing, however, some preliminary remarks of a technical nature are needed.

4.9.1 Minimization with product measures. In each of the models the factorization for the density of $(X_1,\ldots,X_n)$ could be expressed in the form

$$p(x_i;\theta)\ r(x_i;\theta,\phi_i).$$

That is, there were separate partial-remainder factorizations for each observation X_i. In the search for a "nuisance removing distribution," in the sense of solving (58), it makes intuitive sense, based on the discussion in Section 4.5 on additivity, to limit consideration to product measures $Q_{\theta 1}x\ldots xQ_{\theta n}$, where $Q_{\theta i}$ satisfies (58) for $r(x_i;\theta,\phi_i)$. This approach will be taken herein, and because of the focus on the individual observation X_i, the subscript i will often be omitted from the notation.

4.9.4 The families P_θ and P_θ^*. The next point concerns the family P_θ in which the search for $Q_\theta = Q_{\theta i}$ is made. For Models B, C, and E, the underlying measure ν puts mass on a finite set of points. It follows that, since orders of integration can always be interchanged, every probability measure on the space Φ is unbiased. (In none of these models did the support depend on the parameters.)

In Model D. the exponentials with unknown support, it is clear that $(\theta,\phi) << (\theta_0,\theta_0)$ if and only if $\phi<\phi_0$, so the support of Q_θ is limited to the latter set. If lemma 4.4.2 holds, we may search for nuisance eliminating distributions in P_θ^*, which would here include all probability measures with that support. We check the condition of the lemma. For any positive measurable function T,

$$E(T;\theta,\phi)=\int_\phi^\infty \cdots \int_\phi^\infty +\theta^{-k}\exp(k\phi/\theta)\exp(-\Sigma x_i/\theta)dx_1 \cdots dx_n.$$

This is continuous in ϕ by the continuity of the integral and $\exp(k\phi/\theta)$.

Model A can be discussed in the more general context of exponential families. If T is positive and integrable for all (θ,ϕ) and $f(x;\theta,\phi) = c(\theta,\phi)\exp(\Sigma\lambda_1(\theta,\phi)\, S_i(x))$. then the natural parameter space for measure ν

$$N_\nu = \{ (\theta,\phi) ; \int \exp(\Sigma\lambda_i(\theta,\phi)S_i)\, d\nu < \infty \}$$

is contained in N_μ, the natural parameter space for the measure μ defined by $d\mu = T\nu d$, by the integrability assumption. By Theorem 9 of Lehmann ({1959}, p. 52) it follows that both

$$c^{-1}(\theta,\phi) = \int \exp\ (\Sigma\lambda_i S_i)d\nu$$

and

$$c^{-1}(\theta,\phi)\ E(T;\theta,\phi) = \int \exp\ (\Sigma\lambda_i S_i)d\mu$$

are analytic function of the λ_i. Hence under the continuity of the $\lambda_i(\theta,\phi)$ as functions of ϕ, lemma 4.4.2 is satisfied. Model A is such an exponential family model.

4.9.3 Regularity of L_p. Another regularity problem concerns the satisfaction of (54). Once again this is satisfied in Models B, C, and E because the underlying measure ν has finite point support. More generally, if the partial likelihood is in the exponential family, say

$$p(x;\theta) = c(\theta)\ \exp\ (\Sigma\lambda_i(\theta)\ S_i),$$

then the squared partial likelihood ratio is

$$L_p^2 = \exp\ \{\Sigma 2S_i(\lambda_i(\theta) - \lambda_i(\theta_0))\}\ \ c^2(\theta)/c^2(\theta_0).$$

Let $d\mu(x) = f(x;\theta_0,\phi_0)\, d\nu(x)$. Then if EL_p^2 exists, the parameters $2(\lambda_i(\theta) - \lambda_i(\theta_0))$ are in the natural parameter space with respect to μ, so we may appeal again to Lehmann's Theorem 8 to obtain, provided the $\lambda_i(\theta)$ are themselves twice differentiable in θ,

$$(62) \qquad \partial\, E\, L_p^2 \,/\, \partial\theta = 2\, E\, L_p L_p'$$

and

$$(63) \qquad \partial^2 \quad E\, L_p^2 \,/\, \partial\theta^2 = EL_p L_p'' + EL_p'^2.$$

A similar argument to the above, with L_p substituted for L_p^2, yields

$$(64) \qquad \partial\, E\, L_p \,/\, \partial\theta = E\, L_p' = 0$$

and

$$(65) \qquad \partial\, E\, L_p \,/\, \partial\theta^2 = E\, L_p'' = 0,$$

since $E\, L_p = 1$. Evaluating (62) and (63) at $\theta=\theta_0$ gives $\partial\, EL_p^2 \,/\partial\theta \,|_{\theta=\theta_0} = 0$ and $\partial\, EL_p^2 \,/\, \partial\theta^2 \,|_{\theta=\theta_0} = 2I_p$. It follows by L'Hopital's rule that

$$(66) \qquad (EL_p^2 - 1)/(\theta-\theta_0)^2 \to I_p \text{ as } \theta\to\theta_0$$

as desired. The partial likelihoods in Models A and D are exponential family and so satisfy (66).

4.9.4 Computation of I_p. If p is a conditional likelihood, there is a trick which sometimes simplifies the computation of the information I_p. If $E\, L_p'' = E\, L_f'' = 0$, as in the exponential family (see (65)), then the (1,1) entry of the full Fisher matrix is

$$I_{11} = E\,(\partial \ln f/\partial\theta)^2 = E(f'/f)^2 = E((f'/f)^2 - L_f'')$$

$$= E(-\partial^2 \ln f/\partial\,\theta^2)$$

and in the same manner

$$I_p = E(-\partial^2 \ln p/\partial\theta^2).$$

The partial information I_p may therefore be computed from

$$I_{11} = E(-\partial \ln pr/\partial\theta^2) = E(-\partial^2 (\ln p + \ln r)/\partial\theta^2)$$

$$= I_p + E\,(-\partial^2 \ln r/\partial\theta^2).$$

This technique will be used in Model C.

We now discuss the key models in terms of the simple nuisance property and the resulting partial information. Model A has already been presented in 4.4.6 and 4.4.7. Equations (33) and (34) demonstrate that the remainder

likelihood, the marginal distribution of the sample means, has for nuisance eliminating distribution a product of normal measures (31). The remaining models will be presented in increasing order of difficulty.

4.9.5 Model E. For each i the observations are a pair of independent Bernoulli random variables (X_i, Y_i), each with probability of success $\frac{1}{2} + \theta\phi_i$. The parameter θ is in the interval $(0, \frac{1}{2})$ and ϕ_i is +1 or -1. We let $S_i = X_i + Y_i$ and $W_i = I(S_i = 1)$, where I is the indicator function. In Section 2.8.4 we reduced to $W_1, \ldots, W_n$ by an invariance argument. In this case then, the partial likelihood p for each i is the marginal for W. The remainder is the conditional likelihood of S given W.

It is now shown that the remainder is a simple nuisance. If the conditioning event is W=1, then S=1 with probability one. If W=0, then

$$P(S=0 \mid W=0) = (\tfrac{1}{2}+\theta\phi)^2/((\tfrac{1}{2}+\theta\phi)^2 + (\tfrac{1}{2}-\theta\phi)^2)$$

$$P(S=1 \mid W=0) = 0$$

$$P(S=2 \mid W=0) = 1-P(S=0 \mid W=0)$$

Condition (58) will be shown for null $(\theta_0, 1)$. The case $(\theta_0, -1)$ differs only in details. We exclude null $\theta_0 = 0$.

When the conditioning event is W=1, then any mixing distribution satisfies (58), so we may restrict attention to showing that S given W=0, which has a Bernoulli distribution, has the same success parameter under (θ, Q_θ) as under $(\theta_0, 1)$. Let $q = Q(1)$ for any probability measure Q on (-1,+1). Let

$$a(\theta) := (\tfrac{1}{2}+\theta)^2/((\tfrac{1}{2}-\theta)^2 + (\tfrac{1}{2}-\theta)^2).$$

The equation $P(S=0 \mid W=0;\theta,Q) = P(S=0 \mid W=0;\ \theta_0\ 1)$ yields by substitution

$$q\ a(\theta) + (1-q)\ (1-a(\theta)) = a(\theta_0).$$

Solving for q gives

$$q = (a(\theta_0) + a(\theta) - 1)/(2a(\theta) - 1). \tag{68}$$

For Q to define a measure Q, it must lie in the interval from 0 to 1. Since $a(\theta)>\frac{1}{2}$ for $\theta\neq 0$, the right hand side of (68) is positive. Also the right hand side of (68) is less than 1 if and only if $a(\theta_0)<a(\theta)$. The function $a(\theta)$ is monotone increasing in $(0,\frac{1}{2})$, so that $a(\theta_0)<a(\theta)$ is implied by $\theta_0<\theta$. It follows that for $\theta\to\theta_0^+$, the Q_θ distribution defined by (68) is nuisance eliminating in the

sense of (58).

We now compute the partial information so as to compare it with the J-information. The Fisher's information I_p in the marginal distribution, since W is just a Bernoulli $(\frac{1}{2}-2\theta^2)$ random variable, is

$$I_p = 16\theta^2/(\tfrac{1}{2}+2\theta^2)(\tfrac{1}{2}-2\theta^2)$$

Since the J-information was, from 4.2.2, $2(\frac{1}{4}-\theta^2)^{-1}$, the ratio is

$$I_p/J = 2\theta^2/(\tfrac{1}{4}+\theta^2) = 1 - \{(\tfrac{1}{2}-2\theta^2)/(\tfrac{1}{2}+2\theta^2)\},$$

a function which is monotonely increasing from 0 to 1 as θ goes from 0 to $\frac{1}{2}$. Since the factorization is complete, this is also the ratio of the K-information to the J-information.

4.9.6 Model B. For each i we observe a Bernoulli pair (X_i,Y_i) whose success parameters have a common log odds ratio θ over i. It has been argued (Section 2.6) that a natural partial likelihood to use is the conditional likelihood of X given X + Y. The remainder likelihood is thus the marginal for X+Y. If we let $\rho=(1+e^{\theta+\phi})^{-1}(1+e^{\phi})^{-1}$, then the distribution may be expressed

$$r(0;\theta,\phi) = \rho,$$

$$r(1;\theta,\phi) = (1+e^{\theta})\, e^{\phi}\rho,$$

and $$r(2;\theta,\phi) = e^{\theta+2\phi}\rho.$$

Because any two probabilities determine the third, it is sufficient to find a measure Q such that $r(x;\theta,Q)=r(x;\theta_0,\phi_0)$ for x=0 and 1.

Define $\alpha_\theta(\phi) = r(0;\theta,\phi)$ and $\beta_\theta(\phi) = r(1;\theta,\phi)$. Then as ϕ goes from $-\infty$ to $+\infty$, $(\alpha_\theta(\phi),\beta_\theta(\phi))$ traces out a path in the unit square. The following items can be shown:

(i) $(\alpha_\theta(\phi),\beta_\theta(\phi)) \to (1,0)$ as $\phi\to-\infty$.

(ii) $(\alpha_\theta(\phi),\beta_\theta(\phi)) \to (0,0)$ as $\phi\to+\infty$.

(iii) The paths for θ and $-\theta$ overlap. No other path, however, intersects the path for θ on the interior of the square.

(iv) The path is continuous. For each value α in (0,1), there is a unique ϕ_α in R such that $\alpha_\theta(\phi_\alpha) = \alpha$. That is, there is an invertible onto function between $(-\infty,+\infty)$ and (0,1) which takes ϕ to α. Write $\beta_\theta(\alpha):=\beta_\theta(\phi_\alpha)$.

(v) As a function of α, $\beta_\theta(\alpha)$ is continuous and differentiable. It is concave, and has maximum value at $\phi_\alpha = -\theta/2$, where

$$\sup_\alpha \beta_\theta(\alpha) = (1+2/(e^{-\theta/2}+e^{\theta/2}))^{-1}. \tag{69}$$

This supremum is increasing in $|\theta|$.

From (i) through (v) we can draw the following conclusions. If $|\theta|>|\theta_0|$, then $\sup_\alpha \beta_\theta(\alpha)>\sup_\alpha \beta_{\theta_0}(\alpha)$. Because the paths do not cross, and they are continuous, $\beta_\theta(\alpha)>\beta_{\theta_0}(\alpha)$ for all α. Let $\alpha_0=\alpha_{\theta_0}(\phi_0)$. By continuity, on the interval $(0,\alpha_0)$, $\beta_\theta(\alpha)$ takes on all values in $(0,\beta_\theta(\alpha))$. Hence there exists $\alpha_1<\alpha_0$ such that $\beta_\theta(\alpha_1)=\beta_{\theta_0}(\alpha_0)$. Similarily, there is $\alpha_2>\alpha_0$ such that $\beta_\theta(\alpha_2) \cong \beta_{\theta 0}(\alpha_0)$. Thus there exist $\phi_1 := \phi_{\alpha_1}$ and $\phi_2 := \phi_{\alpha_2}$ such that $\beta_\theta(\phi_1) = \beta_\theta(\phi_2) = \beta_{\theta 0}(\phi_0)$. That is,

$$r(1;\theta,\phi_1) = r(1;\theta,\phi_2) = r(1;\theta_0,\phi_0). \tag{70}$$

Also, because $\alpha_1<\alpha_0<\alpha_2$, there exists q in (0,1) such that $q\alpha_1 + (1-q)\alpha_2 = \alpha_0$. That is,

$$q\, r(0:\theta,\phi_1) + (1-q)\, r(0;\theta,\phi_2) = r(0;\theta_0,\phi_0). \tag{71}$$

By (70) and (71), the measure Q_θ with mass q at ϕ_1 and mass 1-q at ϕ_2 yields $r(x;\theta,Q_\theta) = r(x;\theta_0,\phi_0)$ for x = 0,1, and 2, provided $|\theta|>|\theta_0|$.

The computation of the Fisher's information in the partial likelihood will be performed in Section 4.9.9, when Model C is discussed.

4.9.7 Mathematical Interlude. As seen in 4.9.6, determining if a likelihood is a simple nuisance can be quite difficult even in a simple model. Better techniques are needed for determining if the family of equations (58) have a solution Q_θ. The following theory has proved somewhat fruitful.

Suppose that (58) can be reduced a condition on the moments of Q. There is an extensive theory developed about whether a distribution Q exists on R with any given moment sequence $\mu_0, \mu_1, \ldots$, where $\mu_k = \int x^k \, dQ(x)$. The reader is referred to *The Laplace Transform* by David Widder (1941). In the models remaining, the following results from that source are used. Define the k-th moment matrix by

$$M_k = \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{k/2} \\ \mu_1 & \mu_2 & \cdots & \mu_{k/2+1} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \mu_{k/2} & \cdots & & \mu_k \end{bmatrix} \quad \text{if k is even,}$$

and

$$M_k = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_{(k+1)/2} \\ \mu_2 & \mu_3 & \cdots & \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \mu_{(k+1)/2} & \cdots & & \mu_k \end{bmatrix} \quad \text{if k is odd.}$$

4.9.7.1 Theorem (Widder {1941}, p. 138) A necessary and

sufficient condition that the equations

$$(71) \qquad \int_0^\infty t^n \, dP(t) = \mu_n$$

have a nondecreasing solution P with infinitely many points of increase is that det $M_k > 0$, for k=0,1,2,....

Notice that this condition is equivalent to the matrices M_i all being positive definite, as a symmetric matrix is positive definite if and only if all its principal leading minors have positive determinants (Seber {1971}, p. 35).

4.9.7.2 Theorem. (Widder {1941}, p. 138) A necessary and sufficient condition that the equations (71) have a nondecreasing solution with finitely many points of increase is that the matrices M_k be nonnegative definite for all k, with at least one positive semidefinite.

These results are now used in Model D.

4.9.8 Model D. In the transformed problem (see Section 2.6.3) $Z_1^*, Z_2, \ldots, Z_k$ are independent random variables, where each of $Z_1^*+k\phi, Z_2, \ldots, Z_k$ have an exponential distribution with mean θ. Let $Y=Z_1^*$. It was argued, based both on invariance and an Andersen-conditioning argument that

a reasonable partial likelihood was the marginal likelihood of $Z_2,\ldots,Z_k$. Hence the remainder likelihood is the marginal for Y.

In showing that there exists a nuisance eliminating distribution, we may assume without loss of generality that $\phi_0=0$, as other cases can be translated into that one. The requirement that $(\theta,\phi)<<(\theta_0,\phi_0)$ then implies that the measures in P^*_θ must have nonnegative support. Under (θ,ϕ), $Y \overset{L}{=} \theta W + n\phi$, where W is a unit exponential. Let U/n be a random variable with distribution Q, independent of W. Then under (θ,Q),

$$Y \overset{L}{=} \theta W + U.$$

Hence the equation

$$\theta W + U \overset{L}{=} \theta_0 W \tag{72}$$

expresses the likelihood relationship (58). How do we establish (72)? Suppose that there exists a distribution Q for U such that the moments of $\theta W + U$ are the same as those of $\theta_0 W$. Equation (72) then follows if those moments uniquely determine the distribution. The following criterion for the latter is from Breiman ({1968}, p. 182).

4.9.8.1 Proposition. If

$$\limsup_{n \to \infty} \mu_n^{-1}/n < \infty,$$

then there exists at most one distribution function F satisfying

$$\mu_n = \int x^k \, dF(x).$$

The moments of $\theta_0 W$ are $\mu_n = \theta_0^n n!$. Proposition 4.9.8.1 holds because

$$\lim_{n \to \infty} (\theta_0^n \, n!)^{n^{-1}} /n = \theta_0/e$$

by Sterlings formula (Rao {1973}, p. 59).

Hence we search for a distribution Q such that the moments of $\theta W + U$ match those of $\theta_0 W$. Let the k-th moment of U be b_k. Setting the moments of $\theta W + U$ and $\theta_0 W$ equal gives the system of equations:

$$\sum_{k=0}^{n} \binom{n}{k} k! \, \theta^k b_{n-k} = \theta_0^n \, n! \quad \text{for } n = 1,2,3,\ldots.$$

Put into matrix form, this becomes

$$A_\theta \beta = \alpha$$

where $\beta^t = (1, b_1, b_2, \ldots, b_n)$, $\alpha^t = (1, \theta_0 1!, \theta_0^2 2!, \ldots, \theta_0^n n!)$, and A_θ is the lower triangular (n+1) by (n+1) matrix defined by

$$(A_\theta)_{ij} = (i-1)!/(j-1)!\ \theta^{i-j} \quad \text{for } i \geq j$$

$$= 0 \quad \text{else.}$$

The matrix $(A_\theta)^{-1}$ is simple in structure, consisting of 1's down the diagonal, positive entries in positions (i-1,i), and zeroes elsewhere:

$$(A_\theta^{-1})_{ii} = 1 \quad , \quad i = 0,1,\ldots,n,$$

$$(A_\theta^{-1})_{i-1,i} = -i\theta \quad , \quad i = 1,2,\ldots,n,$$

and $$(A_\theta^{-1})_{i,j} = 0 \quad \text{else.}$$

This simple inversion works because of the relationship:

$$(-m\theta,\ 1)\ ((m-1)!/k!\ \theta^{m-1-k},\ m!/k!\ \theta^{m-k})^t = 0.$$

Solving the equation for β yields

$$\beta = A_\theta^{-1}\alpha = (1, (\theta_0-\theta)1!\theta_0^{1-1}, \ldots, (\theta_0-\theta)n!\theta_0^{n-1})^t. \tag{73}$$

Can (73) be a moment vector? First notice that θ must be less than θ_0, as the positive random variable U must have positive moments. Assuming this is true, notice that

$$\theta_0(\theta_0-\theta)^{-1}\beta^t = (\theta/(\theta_0-\theta),0,0,\ldots,0) + (1,1!\theta_0,2!\theta_0^2,\ldots,\theta_0^n n!),$$

where the last vector is the moment sequence for $\theta_0 W$. Let M_k be the moment matrix of $\theta_0 W$, let M_k^* be the moment matrix for $\theta W+U$, and let D_k be the matrix of the same dimension as M_k with $\theta/(\theta_0-\theta)$ in the (1,1) position and zeroes elsewhere. Then

$$M_k^* = M_k\,(\theta_0-\theta)/\theta \quad \text{for k odd}$$

and

$$M_k^* = M_k\,(\theta_0-\theta)/\theta + D_k \quad \text{for k even.}$$

For k odd, M_k^* is positive definite because M_k is, by Theorem 4.9.7.1. For k even, M_k^* is positive definite because M_k is positive definite and D_k is nonnegative definite. The conclusion is that for $\theta<\theta_0$, equation (72) has a solution and hence so does equation (58). Thus the marginal likelihood of Y is a simple nuisance.

Recall from 4.2.8 that the J-information was k/θ^2. For this problem, the partial likelihood is the density

for the k-1 independent exponential (θ) random variables $Z_2,\ldots,Z_n$. The Fisher's information in p, which is also the K-information, is $I_p(\theta;\phi) = (k-1)/\theta^2$. Comparing the two informations, one sees that the adjustment for the nuisance parameter is the equivalent of the loss of one observation.

4.9.9 Model C. The observations are pairs of independent binomials (X_i,Y_i), with sample sizes r_i,s_i respectively. The parameter of interest θ is the log odds ratio, which is constant through all observations. The partial likelihood suggested by the Andersen conditioning argument, and earlier by Fisher, is the conditional likelihood of the pair given the sum X+Y. The remainder likelihood is therefore that of the sum X+Y.

This model is the most tantalizing of the models presented here. The conditional argument is both frequently used and somewhat controversial. Unfortunately, as will be seen, some of the uncertainty about the optimality of the CMLE will still remain after tha analysis of this paper. First, we check to see if the simple nuisance criteria works for r and s greater than one (recall that it worked for r=1 and s=1 in Model B). It turns out that it doesn't for the Type I likelihood, but that for a large class of null distributions Q_0, it works in

the Type II likelihood. It is conjectured that this is a model in which the factorization is complete, but the simple nuisance criterion fails.

Define the function $f_k(\theta)$ to be the coefficient of $e^{\phi k}$ in the polynomial expansion of $(1+e^{\theta+\phi})^r\ (1+e^{\phi})^s$ in powers of e^{ϕ}. Define

$$\rho(\theta,\phi) = P(X+Y = 0\ ;\ \theta,\phi)$$

$$= (1+e^{\theta+\phi})^{-r}\ (1+e^{\phi})^{-s}.$$

Then the marginal distribution of X + Y may be expressed as

$$P(X+Y = k\ ;\ \theta,\phi) = \rho(\theta,\phi)e^{k\phi}\ f_k(\theta). \tag{74}$$

Thus in this model equation (58) is equivalent to the followint r+s+1 equations:

$$\int \rho(\theta,\phi)\ e^{k\phi} f_k(\theta)\ dQ_\theta = \rho(\theta_0,\phi_0)\ \exp(k\phi_0) f_k(\theta_0) \tag{75}$$

$$\text{for } k = 0,1,\ldots,r+s.$$

Let $\mu_k(\theta) = f_k(\theta_0)/f_k(\theta)$. Suppose that there exists a measure w on R such that

$$(76) \qquad \int e^{k(\phi-\phi_0)}\, dw(\phi) = \mu_k(\theta) \quad k = 0,1,2,\ldots,r+s.$$

Then $dQ(\phi) = p(\theta_0,\phi_0)/p_0(\theta,\phi)\, dw(\phi)$ would be a solution to equation (75). The measure Q so defined would of necessity be a probability measure, because summing the equations (75) over k gives $\int 1\, dQ = 1$. Now (76) holds only if there exists a measure w* on $(0,\infty)$ such that

$$(77) \qquad \int x^k\, dw^*(x) = \mu_k(\theta) \quad k = 0,1,2,\ldots,r+s,$$

so we may apply the moment theory of Section 4.9.7. An important result is stated in the following lemma.

4.9.9.1 Lemma. For sample sizes r=s=2 and null $(\theta_0,\phi_0) = (0,\phi_0)$, the marginal likelihood of X+Y is not a simple nuisance.

Proof. It will be shown that det $M_4 < 0$ for the purported moment sequence (77). The matrix is

$$M_4 = \begin{bmatrix} 1 & 2(1+e^{\theta})^{-1} & 6(1+4e^{\theta}+e^{2\theta})^{-1} \\ 2(1+e^{\theta})^{-1} & 6(1+4e^{\theta}+e^{2\theta})^{-1} & 2(1+e^{\theta})^{-1}e^{-\theta} \\ 6(1+4e^{\theta}+e^{2\theta})^{-1} & 2(1+e^{\theta})^{-1}e^{-\theta} & e^{-2\theta} \end{bmatrix}$$

Making the substitutions $\alpha=(1+e^{\theta})^{-1}$ and $\beta=\alpha(1-\alpha)$ the

determinant may be expressed

$$\det M_4 = \frac{\alpha^6 (1-4\beta)^3 (-2)}{\beta^2 (1+2\beta)^3}$$

Since $\beta \leq \frac{1}{4}$, with strict inequality for $\theta \neq 0$, the determinant is strictly negative. It follows that the matrix M_4 cannot be nonnegative definite. Thus the remainder is not a simple nuisance.

4.9.9.2 Theorem. Under the null parameter (θ_0, Q_0), where Q_0 is any distribution on ϕ with infinitely many points of increase, the marginal likelihood of X+Y is a simple nuisance.

Proof. Let $u_k(\theta) = P(X+Y = k ; \theta_0, Q_0)/f_k(\theta)$. Define measure w by $dw(\phi) = \rho(\theta_0, \phi)\, dQ_0(\phi)$. Then w is a finite measure, as $\rho(\theta_0, \phi)$ is a probability and so is between 0 and 1. The distribution function for w has infinitely many points of increase because Q_0 does and $\rho(\theta_0, \phi)$ is strictly positive. Notice that

$$P(X+Y = k; \theta_0, \phi_0) = \int \rho(\theta_0, \phi)\, e^{k\phi} f_k(\theta_0)\, dQ_0$$

implies, by division by $f_k(\theta_0)$, that

$$\mu_k(\theta_0) = \int e^{k\phi}\ \rho(\theta_0,\phi)\ dQ_0$$

$$= \int e^{k\phi}\ dw(\phi) \qquad k = 0,1,2,\ldots,r+s$$

From the last, it is clear that $\mu_k(\theta_0)$ is a moment sequence for a distribution w with infinitely many points of increase. So the sequence of moment matrices $M_0(\theta_0),\ldots,$ $M_{r+s}(\theta_0)$ all have positive determinants. Since the $\mu_k(\theta)$ are continuous functions of θ, so are the corresponding moment matrices $M_k(\theta)$. Hence for θ in a neighborhood of θ_0, det $M_k(\theta)>0$ for $k = 0,1,2,\ldots,r+s$. It follows that $\mu_k(\theta)$ is also a sequence of moments for some distribution, provided θ is sufficiently close to θ_0. Hence there exists a measure w_θ on R such that

$$\int e^{k\phi}\ dw_\theta(\phi) = \mu_k(\theta) \quad \text{for } k = 0,1,2,\ldots,r+s.$$

If we now define Q_θ by $dQ_\theta\ (\phi) = \rho^{-1}(\theta,\phi)\ dw_\theta(\phi)$, then

(78) $$P(X+Y = k;\ \theta,Q_\theta) = \int e^{k\phi}\ f_k(\theta)\ \ \rho(\theta,\phi)\ dQ_\theta$$

$$= \int e^{k\phi}\ f_k(\theta)\ dw_\theta(\phi)$$

$$= f_k(\theta)\ \ \int e^{k\phi}\ dw_\theta(\phi)$$

$$= f_k(\theta)\ \mu_k(\theta)$$

$$= P(\ X+Y = k\ ;\theta_0,Q_0).$$

It is clear by summing over both sides of (78) over k that Q_θ must be a probability measure, and so the simple nuisance property is established.

We now compare the Fisher information about θ in the full density with that in the conditional. The Fisher information in the pair (X_i, Y_i) is

$$I_i(\theta, \phi_i) = \begin{bmatrix} r_i p_i(1-p_i) & r_i p_i(1-p_i) \\ r_i p_i(1-p_i) & r_i p_i(1-p_i)+s_i q_i(1-q_i) \end{bmatrix}$$

where p_i and q_i are the respective success parameters of X_i and Y_i. Hence the Fisher information about θ available in the presence of ϕ is

(79) $$I_i(\theta;\phi_i)=r_i p_i(1-p_i)q_i(1-q_i)/(r_i p_i(1-p_i)+s_i q_i(1-q_i)).$$

The partial information in the conditional likelihood will be computed using (67) of 4.9.4. That is, we will subtract the information in the marginal X+Y from the (1,1) entry of the full Fisher matrix. Writing the marginal density as in (74)

$$\partial^2 \ln r/\partial\theta^2 = \partial(\ln \rho(\theta,\phi)\, e^{k\phi}\, f_k(\theta))/\partial\theta^2$$

$$= - rp(1-p) + \partial^2 \ln f_k(\theta)/\partial\theta^2.$$

Taking expectations on both sides, and applying (67) yields

$$(80) \qquad I_p(\theta;\phi) = E\,(-\partial^2 \ln f_{X+Y}(\theta)/\partial\theta^2).$$

The calculations are now carried out for Model B, where r=s=1. In this case, $f_0(\theta)=1$, $f_1(\theta)=(1+e^{\theta})$ and $f_2(\theta)=e^{\theta}$. So, if $p = P\{x=1\}$,

$$(81) \qquad I_p = P(X+Y = 1;\ \theta,\phi)\ e^{\theta}/(1+e^{\theta})^2$$

$$= p(1-p)(1+e^{\theta+\phi})(1+e^{\theta})^{-1}(1+e^{\phi})^{-1}.$$

Recall that in this case the factorization is complete.

4.10 The symmetry problem. In the Type I model, the following example demonstrates the need for further development of the preceding information theory. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables, where X_i is normally distributed with mean θ and variance ϕ_i. This example was previously used to demonstrate the need to consider carefully the sequence $\phi_1, \phi_2, \ldots$ in constructing an asymptotic model (see 3.2.1). The MLE for θ, when $\phi^{(n)}=(\phi_1,\ldots,\phi_n)$ is assumed known, is

$$(82) \qquad T(\phi^{(n)}) = \sum_{i=1}^{n} X_i \phi_i^{-1} \Big/ \sum_{i=1}^{n} \phi_i^{-1}.$$

If we now consider $T(a^{(n)})$ as a function of the constants $(a_1,\ldots,a_n)$, then for every null $(\theta_0,\phi_0^{(n)})$ the estimator $T(a^{(n)})$ is unbiased. Further, when $a^{(n)}=\phi_0^{(n)}$, it attains the lower bound for the null $(\theta_0,\phi_0^{(n)})$ based on Fisher's information. It follows that the Fisher's bound is the best possible. Thus the K-information must equal the Fisher's information, since otherwise the K-information would provide a better bound.

In this, or any other Type I model, in the face of the unknown character of the sequence $\phi_1,\ldots,\phi_n$, it seems unlikely that one would wish to choose an estimator, such as $T(a^{(n)})$, which was locally optimal, at $\phi_0^{(n)}=a^{(n)}$, but which would perform rather poorly for a permutation of the null sequence.

One way to approach this problem is to use as a loss function for the unbiased estimator T the function

$$(83) \qquad (n!)^{-1} \sum_{\pi} E\{(T-\theta)^2 \; ; \; \theta,\pi\phi^{(n)}\}$$

where $\pi\phi^{(n)}$ ranges over the permutations of the sequence $\phi_1,\ldots,\phi_n$. Under such a loss it can be shown that an asymmetric estimator is no longer optimal: The loss function (83) is equivalent to computing the variance of T with respect to the density

$$(84) \qquad f^*(x;\theta,\phi^{(n)}) = (n!)^{-1} \sum_{\pi} f(x;\theta, \pi\phi^{(n)}).$$

In this density the order statistics $0^{(n)}$ are sufficient. Let E* be expectation with respect to (84). For any unbiased estimator T, the estimator $E^*(T|0^{(n)})$ is unbiased, is symmetric, and by the Rao-Blackwell Theorem (see Bickel and Doksum {1977}, p. 121) has variance at least as small as T. In fact, under completeness of $0^{(n)}$, $E^*(T|0^{(n)})$ is the unique minimum variance unbiased estimator, provided it has finite variance (see Bickel and Doksum {1977}, p. 122). Notice that in the normal example, $E^*(T(a^{(n)})|0^{(n)}) = \bar{X}$.

What happens if one uses density f* in lieu of f in the preceding theory of unbiased estimation? Does one alter the corrected Fisher's information? In the normal example, the conditional distribution of $X^{(n)}$ given $0^{(n)}$ does not depend on θ, therefore $0^{(n)}$ contains all the corrected Fisher information. Does one alter the K-information? This has not yet been worked out. It is worth pointing out, however, that a lower bound based on the K-information in the order statistics cannot be less than the K-lower bound based on the full likelihood. Hence if there is a symmetric estimator which meets the bound, as in Models A, D, and E, the K-information in the marginal must be the same as in the full likelihood.

4.11 Chapter summary. The chapter started with the notion of J-information, which proved to be a close relative of the Fisher information. It had the added benefit of being well defined when the nuisance parameter is not real valued or when differentiability requirements do not hold.

The K-information, based on Barankin's lower bound, went a step beyond the J-information. It was the key measure used herein to check the optimality of the partial likelihood factorization.

The computation of the K-information, however, is by no means simple. In this chapter the main tool used was the notion of the simple nuisance likelihood. It was used to determine if a factorization was complete in the sense that all the information about θ necessary for unbiased estimation could be presumed to be in the partial likelihood.

It was found that in the Type I models, save Model C, the remainder likelihoods were simple nuisances. In all the Type II models, with a restriction on Model C's null nuisance parameter, the remainder likelihoods were simple nuisances. In Models A, D, and E these results were compelling, as the resulting lower bounds were tight. In Models B and C, the results were merely suggestive, as no unbiased estimators exist.

The next chapter applies the insights into information gained in this chapter to the problem of lower bounds for the asymptotic variances of asymptotically normal estimates.

CHAPTER V

Lower Bounds for Asymptotically Normal Estimates

5.1 Chapter Introduction. In Chapter 4 lower bounds for unbiased estimation were developed which had the advantages of being both well-defined in cases where Φ was not Euclidean and more powerful than the Cramer-Rao bound when Φ was Euclidean. This chapter extends these results to lower bounds for consistent asymptotically normal estimates by using the same basic tool, mixing distributions over the nuisance parameter space.

The methods used herein are modifications and extensions of those of Bahadur (1964). The necessary groundwork of creating an asymptotic model suitable for those methods is made in Sections 5.2 through 5.5. The K-information, is, unfortunately, not directly suitable for this model, as certain smoothness assumptions must be satisfied by the mixing distributions. The effect of these assumptions is discussed in Section 2.9.

A lower bound theorem for the asymptotic variance of asymptotically normal estimates is developed in Sections 5.6 through 5.8. Section 5.10 discusses circumstances in which the criterion for estimators of the lower bound theorem is so strict that there are no such estimates. A solution to this problem is developed in 5.11 and 5.12.

5.2 The asymptotic model. The sequence $X_1, X_2, \ldots$ of random variables is assumed to be so distributed that the marginal distribution of $X^{(n)}=(X_1,\ldots,X_n)$ is generated by the density

$$f^{(n)}(X^{(n)};\ \theta,\phi)$$

with respect to a measure $\nu^{(n)}$ on the observation space $\{(x_1,\ldots,x_n)\}$. It is further supposed that the density $f^{(n)}$ has a factorization into component densities f_i:

$$(1) \qquad f^{(n)}(x^{(n)};\ \theta,\phi) = \prod_{i=1}^{n} f_i(x^{(i)};\ \theta,\phi).$$

The functions f_i are not functions of the index n. The model (1) will be called the general model. The parameter θ is assumed to be real valued, from open set Θ. The parameter ϕ is assumed to be from parameter space Φ.

The Type I model fits the general model, with the interpretation that the nuisance parameter space is Φ^{∞}, the space of infinite sequences $\phi^{(\infty)}=(\phi_1,\phi_2,\ldots)$. The density f_i in (1) then becomes the density of observation X_i. For the Type I model, F_n will denote the empirical measure on Φ which puts mass 1/n at each of $\phi_1,\ldots,\phi_n$. An assumption which will be used for some models in the following text is that F_n converges weakly to a probability measure F.

The Type II model fits the general model by the identification $f_i(x^{(i)};\ \theta,\phi) = f(x_i;\theta,Q)$. The nuisance parameter space is P, the space of probability measures on Φ.

Within this framework, attention will be restricted to estimation of θ, rather than a function thereof. Further, in parallel with classical theory, we consider a specialized class of estimators.

5.2.1 Definition. A sequence of measurable functions $T_n(x^{(n)})$ will be called a consistently asymptotically normal (CAN) estimator at (θ,ϕ) provided that there exists a normalizing constant $v=v(\theta,\phi)$, called the asymptotic variance, such that

$$n^{\frac{1}{2}}(T_n-\theta)/v^{\frac{1}{2}} \xrightarrow{L} N(0,1) \quad \text{as } n\to\infty.$$

The essential focus here is on CAN estimators. As pointed out in 1.4.1, however, the problem of superefficiency prevents a useful lower bound theory for the asymptotic variances of CAN estimators. One solution in nuisance-free models was to require the estimators to be uniformly asymptotically normal (CUAN, 1.4.3). The uniformity was required over compact sets of Θ. In his discussion of conditional maximum likelihood estimates

in the Type I model, E.B. Andersen (1973) used the concept of ϕ-uniform convergence (see Section 3.7.1), for which uniformity was required only over the nuisance space. Herein, a weaker requirement will be used.

5.2.2 Definition. The sequence of estimators T_n will be called ϕ-uniformly median unbiased (UMU) at θ if

$$\sup_{\Phi} \{|P(T_n<\theta;\theta,\phi) - \tfrac{1}{2}|\} \to 0 \text{ as } n\to\infty.$$

The requirement that an estimator be UMU is essentially a technical one in order to obviate superefficient estimates, rather than expressing a desirable property of the estimator. The requirement that T_n be ϕ-uniformly normal, as in Andersen (see 3.7.1), could be justified however based on the confidence interval argument of Section 1.4.4. The weaker requirement will be used here because it suffices for the proofs.

5.3 Bahadur's proof. The lower bound of this chapter is derived by making suitable modifications in a proof, given by Bahadur (1964), of the Fisher lower bound for asymptotic variance in a nuisance-free i.i.d. model. His approach is sketched here.

Suppose that $X_1, X_2, \ldots$ are i.i.d. observations from

density $f(x;\theta)$, for θ real valued. The density is assumed differentiable in θ. The score random variables are

$$U_i = \partial \ln f(X_i;\theta) / \partial\theta \ , \ i=1,2,3,\ldots.$$

Then, under regularity assumptions, $n^{-\frac{1}{2}}\Sigma U_i$ is asymptotically normal, with mean 0 and variance $I := E(U_i^2)$.

Rather than consider the scores directly, let $\theta_n = \theta_0 + n^{-\frac{1}{2}}$, let $L(X_i;\theta_n) = f(X_i;\theta_n)/f(X_i;\theta_0)$, and consider

$$(2) \qquad V_n := \sum_{i=1}^{n} \ln L(X_i;\theta_n).$$

Then a Taylor's expansion of V_n yields

$$(3) \qquad V_n \sim n^{-\frac{1}{2}}\Sigma U_f + \tfrac{1}{2}n^{-1}\Sigma U_i'$$

$$\xrightarrow{L} N(0,I) + \tfrac{1}{2}I \quad \text{under } \theta_0, \text{ as } n\to\infty,$$

where $I = I(\theta_0)$. Bahadur also shows that

$$(4) \qquad (V_n - \tfrac{1}{2}I)/\ I^{\frac{1}{2}} \xrightarrow{L} N(I^{\frac{1}{2}},1) \quad \text{as } n\to\infty$$

under laws $P(\ .\ ;\theta_n)$ for V_n.

Each CAN estimate T_n may be turned into a test of $H_0:\theta=\theta_0$ versus alternative $H_n:\theta=\theta_n$. Since V_n generates the most powerful test of this hypothesis (Neyman-Pearson

lemma), the two tests may be compared. Bahadur does this in such a fashion, using (3) and (4), so as to show that for almost all θ (Lebesgue measure), the asymptotic variance of T_n exceeds I^{-1}.

5.4 Directional Scores. We now seek to generalize Bahadur's argument to the nuisance parameter model, starting with an expanded notion of the score function U_i.

Suppose that $f(x;\theta,\phi)$, with ϕ in R^s, satisfies the necessary regularity conditions in θ and ϕ so that the full Fisher's information matrix I exists. The score functions are $U_\theta = \partial \ln f/\partial\theta$ and $U_i = \partial \ln f/\partial\phi_i$. Fix null (θ_0,ϕ_0). For the row vector $w = (w_1,\ldots,w_s)$ in R^s, let $\phi_w(\theta) = \phi_0+(\theta-\theta_0)w$. Then $\phi_w(\theta_0)=\phi_0$, and

$$(5) \qquad U_w := \partial \ln f(x;\theta,\phi_w(\theta))/\partial\theta$$

$$= U_\theta + w_1U_{\phi_1} + \ldots + w_sU_{\phi_s}.$$

This score, which depends on a direction w in R^s, has mean zero and variance $(1,w)I(1,w)^t$. It will be called the direction w score. By lemma 4.2.3, the variance is minimized for direction $w^*=I_{12}I_{22}^{-1}$, at which point Var $U_{w^*}=I_{11}-I_{12}I_{22}^{-1}I_{21}$, the Fisher information corrected for the nuisance parameters. The direction w^* is, in this sense, the least favorable direction.

The scores may be represented in a way which suggests a link to the material of the previous chapter. Let $L(\theta,\phi) = f(X;\theta,\phi)/f(X;\theta_0,\phi_0)$. Then, since $\ln f(x;\theta_0,\phi_0)$ does not depend on θ or ϕ, we may write the score U_θ as $\partial \ln L(\theta,\phi)/\partial\theta\big|_{(\theta_0,\phi_0)}$. Instead of directions w in Φ, consider "directions" in P, the space of probability measures on Φ, defined by functionals $H_\theta:\Theta\to P$ which assign probability measures to points in θ. For our purposes the most important example occurs when H_θ is nuisance eliminating for the likelihood factorization pr. In this case the "direction H" score would be

$$(6) \qquad \partial \ln L(\theta,H_\theta)/\partial\theta = \partial \ln L_p(\theta)/\partial\theta = U_p,$$

the partial score. It is important to note here that H_θ need only exist for $\theta\to\theta_0^+$ or $\theta\to\theta_0^-$, so that the partial in (6) is technically only a left or right derivative. We now formalize the notion of a directional score in P.

5.4.1 Definition. Let H_θ be a function which, for some $c>0$, assigns to each θ in either (θ_0,θ_0+c) or (θ_0-c,θ_0) a probability measure, denoted H_θ, in P. The density $f_H(\theta):=f(x;\theta,H_\theta)$ will be called the direction-H density. The direction-H log likelihood ratio is defined to be

$$D_H(\theta) = \ln L(\theta, H_\theta) = \ln f_H(\theta)/f(x;\theta_0,\phi_0).$$

5.4.2 Definition. If $f_H(\theta) \to f_H(\theta_0)$ as $\theta \to \theta_0^{+(-)}$ and the right (left) derivative

$$U_H(\theta_0) := \lim_{\theta \to \theta_0^{+(-)}} \{D_H(\theta) - D_H(\theta_0)\} \; / \; (\theta - \theta_0)$$

exists, then that derivative is called the direction-H score. The function H_θ will then be called a score producing functional. Define the Fisher's information in direction-H to be

$$I_H(\theta_0) := \mathrm{Var}(U_H(\theta_0)).$$

5.4.3 Definition. A score-producing functional H will be called regular if

a. $D_H'(\theta)$ exists in an open interval adjoining θ_0, and $D_H'(\theta) \to U_H(\theta_0)$ as $\theta \to \theta_0^{+(-)}$.

b. $D_H''(\theta)$ exists in an open interval adjoining to θ_0. The right (left) second derivative at θ_0 exists, and is continuous with $D_H''(\theta)$. Denote this right hand second derivative U_H'.

c. The following interchange of limit and integral can be made:

$$E(U_H) = \lim^{+} E\{(f_H(\theta)-f_H(\theta_0)/f_H(\theta_0)(\theta-\theta_0)\}=0.$$

d. $I_H(\theta_0) = E(-U_H')$.

The above definition applies to any measurable parameter space Φ. If Φ is a Euclidean space, then, under the conditions for the existence of the Fisher's information matrix, the direction-H score is identical to the direction w score if H_θ is defined to put mass 1 at $\phi_w(\theta)$. One advantage of considering directions in P is, of course, that under simple nuisance conditions the partial scores may be so generated (see equation (6)).

5.4.4 Application to simple nuisance. If the direction H is the "nuisance eliminating" direction for a partial likelihood factorization f = pr, then the regularity assumptions a through d above are satisfied if

a. U_p exists in a neighborhood of θ_0.

b. $\partial^2 \ln p/\partial\theta^2$ exists and is continuous in a neighborhood of θ_0.

c. $E(U_p) = 0$.

d. $E(p''/p) = 0$, as this implies that $-E(\partial^2 \ln p/\partial\theta^2) = E(p'/p)^2 = I_p$.

These are all satisfied if p is from a smooth exponential

family (see 4.9.3).

5.5 Convergence Conditions on the Scores. In order to create an asymptotically normal variable to correspond with V_n of equation (2), the directions H_θ will be chosen in such a manner that the log likelihood ratio for the full density is a sum of random variables. Under mixing distribution H at stage n, the log likelihood of the general model,

$$\ln \int f^{(n)}(x;\theta,\phi)\, dH = \ln \int \Pi\, f_i(x^{(i)};\theta,\phi)\, dH,$$

does not necessarily decompose into a sum. This difficulty leads to the use of distinct approaches for models of differing structure.

In the Type I model, where there is a distinct parameter ϕ_i for each f_i, using the product measure $H_\theta^{(n)} = H_{1,\theta} \times \ldots \times H_{n,\theta}$ on Φ^n yields a log likelihood ratio

$$D_H^{(n)}(\theta) = \Sigma \ln L_i(\theta, H_{i,\theta})$$

which is a sum of independent random variables.

In any model where the density f_i depends on the entire nuisance parameter Φ, such as in the Type II model, then insertion of a randomizing measure fails to provide

summability. In effect, it induces correlations among the scores. The remaining alternative is to use the analogy of the J-information. Select ϕ_θ in Φ which is least favorable for each θ and let H_θ put mass 1 there. Then

$$D_H^{(n)} = \Sigma \ln L_i(X;\theta,\phi_\theta).$$

In summary it is assumed that H_θ is chosen so that at stage n there is a score $U_H^{(n)} = \Sigma U_i$ obtained by differentiation from $D_H^{(n)}$. For the Type I and Type II models the U_i will be independent. In the Type II model they will be identically distributed. In the Type I model, however, the null distributions of the U_i may depend on ϕ_i. If the H_θ are nuisance eliminating distributions for a factorization $f = pr$, where p is a marginal likelihood, then the scores U_i are i.i.d., as their distribution depends only on θ.

For more general factorizations, $f^{(n)} = \Pi f_i$, it will be assumed that the scores are uncorrelated, so that

$$\text{Var } (U^{(n)}) = \Sigma \text{ Var } (U_i) = \Sigma I_i$$

where I_i is a directional Fisher's information for density f_i.

The convergence assumptions needed for the proof

are now presented. After each assumption will be a discussion of its validity for Models A through E.

5.5.1 Assumption. $E(U^{(n)})^2/n = \Sigma\ I_i/n$ converges to an asymptotic information number I in $(0,\infty)$.

For any model in which the U_i are i.i.d. random variables, this condition will be satisfied because $I_i=I_j=I$ for all i and j. Thus this condition is satisfied for any model of Type II. Further, it holds for Models A, D, and E of Type I, since in these models the partial likelihoods are marginals. For a Type I model where the partial is not a marginal, the following remarks are relevant.

Assume the probability measure function $H_{i,\theta}$ assigned to the i-th likelihood ratio is itself a function of i through ϕ_i only. This would be the case if we were using nuisance eliminating distributions or least favorable distributions. Then the corresponding directional information is $I_H^{(n)}(\theta;\phi^{(n)}) = n \int I_H(\theta;\phi)\ dF_n(\phi)$, where F_n is the empirical measure of $\phi_1,\ldots,\phi_n$ and $I_H(\theta;\phi_i)$ is the direction-H information for observation X_i. In this case sufficient conditions for assumption 5.5.1 to hold are that $I_H(\theta;\phi)$ is a bounded continuous function of ϕ and that $F_n \xrightarrow{w} F$, a probability measure on Φ. In this case

(7) $$I = \int I_H(\theta;\phi)\, dF(\phi)$$

In Model B, the partial information is

$$I_i = P\{X+Y = 1 ; \theta, \phi_i\}\ e^{\theta}/(1+e^{\theta})^2,$$

which is bounded and continuous in ϕ, so the asymptotic information is

$$I = \int P(X+Y=1;\ \theta,\phi)\, dF(\phi)\ e^{\theta}/(1+e^{\theta})^2$$

Model C does not have a simple nuisance factorization as a Type I Model, so it will not be considered here.

5.5.2 Assumption. $n^{-\frac{1}{2}}U^{(n)} \xrightarrow{L} N(0,I)$.

Again, this condition is satisfied for any model of Type II and for Models A, D, E of Type I, by a central limit theorem for the i.i.d. scores U_i. For models in which the scores are not i.i.d., a different justification must be sought. In the case of Model B, we use Liapunov Theorem (e.g., Rao {1973}, p. 127). That is, if we let $b(\theta;\phi_i) = E(|U_i|^3)$, then assumption 5.5.2 holds if

(8) $$\Big(\sum_{i=1}^{n} b(\theta;\phi_i)\Big)^{1/3} \Big/ \Big(\sum_{i=1}^{n} I(\theta;\phi_i)\Big)^{\frac{1}{2}} \to 0 \quad \text{as } n\to\infty.$$

The ratio in equation (8) may be rewritten

$$n^{1/3} \left(\int b(\theta;\phi)\, dF_n\right)^{1/3} \Big/ n^{\frac{1}{2}} \left(\int I(\theta;\phi)\, dF_n\right)^{\frac{1}{2}}.$$

Hence if it is shown that both b and I are bounded continuous functions of ϕ, this ratio converges to zero and hence 5.5.2 holds by Liapunov's Theorem. In Model B the partial scores are

$$U_i = (X_i - e^{\theta}/(1+e^{\theta}))\ \ I\{X_i+Y_i=1\},$$

where I denotes the indicator function. Hence all absolute moments are bounded by 1. Since the point probabilities are continuous in the finite sample space, all expectations are continuous. It follows that assumption 5.5.2 holds if $F_n \xrightarrow{w} F$ as $n\to\infty$.

5.5.3 Assumption. $n^{-1} \sum_{i=1}^{n} U_i' + I \xrightarrow{p} 0.$

Assumption 5.5.3 holds if the U_i' are i.i.d. random variables because of 5.4.1.d and the weak law of large numbers. In Model B,

$$U_i' = e^{\theta}/(1+e^{\theta})^2\ \ I\{X_i+Y_i = 1\},$$

so the requisite condition is, given that $F_n \xrightarrow{w} F$,

$$n^{-1} \Sigma I\{X_i+Y_i = 1\} \xrightarrow{P} \int P(X+Y=1;\ \theta,\phi)\ dF(\phi).$$

By Chebyshev's Theorem (W.L.L.N.) (e.g., Rao {1973}, p. 112) it suffices that

$$n^{-1} \int I(\theta;\phi)\ dF_n \to 0.$$

This holds because $I(\theta;\phi)$ is bounded and continuous as a function of ϕ.

5.5.4 Assumption. For any sequence θ_n^*, $n = 1,2,3,\ldots$, such that θ_n^* is in $(\theta_0, \theta_0+n^{-\frac{1}{2}})$ for each n,

$$n^{-1} (D_H^{(n)''}(\theta_n^*) - U_H') \xrightarrow{P} 0.$$

The following lemma gives a useful sufficient condition for this assumption.

5.5.5 Lemma. (Bahadur {1964}) If the log likelihoods D_i are of the form $h(S_i)$, where the S_i are a sequence of i.i.d random variables and h is a measururable function of s and the parameters, then the following condition suffices for assumption 5.5.4: there exists a measureable integrable function M(x) such that for all x

$$|D_i''(\theta) - U_i'| \leq M(x)$$

Proof. Let

(8) $$A_i(\delta) = \sup \{|D_i''(\theta) - U_i'| : \theta\varepsilon(\theta_0, \theta_0 + \delta)\}.$$

If we let $m(\delta)$ be the mean of $A_i(\delta)$, then the Lebesgue dominated convergence theorem, using $M(x)$, implies that $m(\delta)\to 0$ as $\delta\to 0$. For all n larger than δ^{-1}, and all $x^{(n)}$,

(9) $$n^{-1}\left|\sum_{i=1}^{n} D_i''(\theta_n^*) - U_i'\right| = n^{-1} \Sigma A_i(n^{-1}) = n^{-1}\Sigma A_i(\delta).$$

By the S.L.L.N.,

(10) $$n^{-1} \Sigma A_i(\delta) \to m(\delta).$$

The conclusion follows by letting $\delta\to 0$ in (9).

For the case of non-i.i.d scores, the following corollary is relevant.

5.5.6 Corollary to Proof. If (10) holds for $A_i(\delta)$ defined as in (8), and $m(\delta)\to 0$, then assumption 5.5.4 holds.

In all models considered except for Type I Model B

the partial scores are i.i.d. and the condition of Lemma 5.5.5 is a regularity condition routinely satisfied by the exponential family.

For Model B, the function A_i of (8) is

$$A_i(\delta) = A(\delta) \quad I(X_i+Y_i=1),$$

where

$$A(\delta) = \sup\{|e^{\theta}/(1+e^{\theta})^2 - e^{\theta_0}/(1+e^{\theta_0})^2| : \theta \in (\theta_0, \theta_0+\delta)\}.$$

Since $A(\delta) \to 0$ as $\delta \to 0$, it is clear that 5.5.6 holds.

5.5.7 Summary of results for partial scores. The convergence conditions of 5.5.1 through 5.5.4 are satisfied by the partial likelihood ratios for all ϕ-sequences for Type I Models A, D, and E, because of their i.i.d. nature. For Type I Model B, they are satisfied provided $F_n \xrightarrow{w} F$. The conditions are satisfied for all Type II models because of the i.i.d. nature of the ratios.

5.5.8 Definition. The score producing functional H_θ will be called normal at (θ_0, ϕ_0) if it is regular and Assumptions 5.5.1 through 5.5.4 are satisfied by the directional log likelihood rations $D_H^{(n)}$.

5.6 Asymptotic Normality of $D_H^{(n)}$. Let $\theta_n=\theta_0+n^{-\frac{1}{2}}$. Assume H_θ is normal. Call the distribution generated by (θ_n,H_{θ_n}) on $(X_1,\ldots,X_n)$ the n-th alternative distribution. Denote the probabilities and expectations with respect to this distribution by P_n and E_n unless otherwise stated. The unsubscripted P and E will continue to represent probabilities and expectations with respect to the null distribution (θ_0,ϕ_0) The log likelihood ratio $D_H(\theta_n)$ determines the most powerful size α test of the null (θ_0,ϕ_0) versus the n-th alternative. We now proceed to compute its asymptotic distribution under both the null and the alternative. Let

$$(10) \qquad Z_n = \{D_H^{(n)}(\theta_n) + I/2\} \;/\; I^{\frac{1}{2}}.$$

5.6.1 Theorem. Under convergence assumptions 5.5.1 through 5.5.4, $Z_n \xrightarrow{L} N(0,1)$ under the null.

Proof. (As in Bahadur, {1964}) It follows from the differentiability of the scores and Taylor's Theorem that

$$(11) \qquad D_H^{(n)}(\theta_n) = n^{-\frac{1}{2}}U^{(n)} + \tfrac{1}{2}n^{-1}U^{(n)\prime} + \tfrac{1}{2}n^{-1}(D_n''(\theta_n^*)-U^{(n)\prime}).$$

By comparing the limits of the first, second, and third summands of the right hand side of (11) with assumptions

5.5.2, 5.5.3, and 5.5.4 respectively, the theorem is proved.

5.6.2 Theorem. Let $C_n = \{x^{(n)} : Z_n < z\}$. Then

$$P(C_n) = N(z) + o(1)$$

and

$$P_n(C_n) = N(z - I^{\frac{1}{2}}) + o(1),$$

where N is the distribution function for the standard normal.

Proof. (As in Bahadur {1964}) The first result restates Theorem 5.6.1. For the second, let I_n be the indicator function for set C_n. Then

$$P_n(C_n) = \int I_n \, f^{(n)}(X ; \theta_n, H_{\theta_n}) \, d\nu^{(n)}(x)$$

$$= \int I_n \exp(D_H^{(n)}(\theta_n)) \, f^{(n)}(x; \theta_0, \phi_0) d\nu^{(n)}(x)$$

$$= \int \exp(-I/2) \, I_n \exp(I^{\frac{1}{2}} Z_n) \, dP_{Z_n}(z)$$

where P_{Z_n}, the null distribution of Z_n, converges to N, the standard normal. Since the integrand is a bounded continuous function the last expression may be written

$$= \exp(-I/2) \int_{-\infty}^{z} \exp(I^{\frac{1}{2}} y) \, dN(y) + o(1),$$

which by a change of variable yields

$$\exp(-I/2)\exp(I/2)\,N(z-I^{\frac{1}{2}}) + o(1),$$

the desired result.

5.7 A lower bound theorem. The lower bound theorem in this section provides a sufficient condition for the directional score information to provide a lower bound for the asymptotic variance of an asymptotically normal estimate. In the next section, the sufficient condition will be seen to hold almost universally in the parameter space.

5.7.1 Theorem. If T_n is CAN at null (θ_0,θ_0), with asymptotic variance v, and

$$\liminf_{n\to\infty} P_n(T_n \leq \theta_n) \leq \tfrac{1}{2}, \tag{12}$$

then $v \geq I_H^{-1}(\theta_0;\phi_0)$ for all normal H.

Proof. (Bahadur {1964}) Choose $z>I^{\frac{1}{2}}$. Define C_n as in Theorem 5.6.2. Let A_n be the complement of C_n. By that theorem

$$\lim_n P_n(A_n) < \tfrac{1}{2}$$

By assumption (12), if $B_n = \{x^{(n)} : T_n \geq \theta_n\}$, then

$$\limsup_{n \to \infty} P_n(B_n) \geq \tfrac{1}{2}.$$

Hence there exists a subsequence $m_1, m_2, m_3, \ldots$ of the positive integers such that

$$P_n(B_n) > P_n(A_n) \quad , \; n = m_1, m_2, \ldots . \tag{13}$$

For each $n = m_i$ regard A_n and B_n as critical regions for testing the null against the n-th alternative. For its size, the region A_n provides the test with the greatest power. Equation (13) states that it has less power than B_n. It follows that A_n must have a smaller size for each n in that same subsequence:

$$P(B_n) < P(A_n) \quad , \; n = m_1, m_2, \ldots . \tag{14}$$

The asymptotic normality of T_n yields

$$P(B_n) = P\{T_n \geq \theta_n\}$$

$$= P\{(T_n - \theta_0) n^{\frac{1}{2}} v^{-\frac{1}{2}} \geq v^{-\frac{1}{2}}\} = 1 - N(v^{-\frac{1}{2}}) + o(1).$$

From Theorem 5.6.2, $P(A_n)=1-N(z)+o(1)$, and so (13) implies that

$$1 - N(v^{-\frac{1}{2}}) + o(1) \geq 1 - N(z) + o(1).$$

Hence $v^{-\frac{1}{2}} \leq z$. Since $z > I^{\frac{1}{2}}$ was arbitrary, the theorem is proved.

5.8 The universality of the bound. For each null point (θ_0, ϕ_0) we may attempt to construct a log likelihood ratio $D_H(.)$, where the normal score generating functional H_θ is a function of the null (θ_0, ϕ_0). If no such functional exists for a point (θ_0, ϕ_0), the asymptotic information is defined to be $I^*(\theta_0; \phi_0) = \infty$. Otherwise, the asymptotic information I^* is defined to be infinum of $I_H(\theta_0 : \phi_0)$ over all normal score generating functionals H.

It is thus presumed that at each parameter point there is an asymptotic information I^*. The question we now ask is, given an estimator T_n which is CAN on a subset of $\Theta \times \Phi$, on "how many" points of that subset can the asymptotic variance be smaller than I^{*-1}? The answer, given below, is that I^{*-1} is a lower bound for the asymptotic variance of all T_n UMU for all θ, at all null points (θ_0, ϕ_0), with the possible exception of a Lebesgue measure zero set of θ_0.

To show this, regard θ as a real variable and, for all n and θ, let

$$h_n(\theta) := \sup_{\phi} \{|P(T_n<\theta : \theta,\phi) - \tfrac{1}{2}|\} \text{ if } \theta \text{ in } \Theta,$$

$$= 0 \quad \text{else.}$$

It is clear that $h_n(\theta)$ is between 0 and $\frac{1}{2}$. Under the assumption of UMU, $h_n(\theta)\to 0$ as $n\to\infty$. Now let $g_n(\theta)=h_n(\theta_n)$.

5.8.1 Lemma. (Bahadur {1964}) There exists a set N of Lebesgue measure zero and a sequence $m_1, m_2, \ldots$ of increasing integers such that $\lim_r g_{m_r}(\theta) = 0$ if $\theta \notin N$.

Proof. Letting N be the standard normal distribution,

$$\int g_n(\theta)\, dN(\theta) = \int f_n(\theta+n^{-\frac{1}{2}})\, dN(\theta),$$

which by a change of variable equals

$$\int f_n(\theta) \exp\{-(2n)^{-1} + n^{-\frac{1}{2}}\theta\}\, dN(\theta).$$

Lebesgue's dominated convergence theorem then implies that the latter converges to zero. Hence there exists a sequence m_r such that, except for a N-null set, g_{m_r} goes to zero

as $n \to \infty$. Since N-null sets are Lebesgue null sets, the lemma is proved.

5.8.2 Theorem. If T_n is UMU for all θ, then $I^{*-1}(\theta_0;\phi_0)$ is a lower bound for the asymptotic variance of T_n at all parameter points at which T_n is CAN, with the possible exception of a set of θ_0 with Lebesgue measure zero.

Proof. Lemma 5.8.1 proves that

$$\sup_{\phi} \left|P(T_n<\theta_n;\ \theta_n,\phi) - \tfrac{1}{2}\right| \to 0 \text{ as } n=m_r\to\infty,$$

for θ_0 not in a null set N. It follows that for any family of distributions H_{θ_n} (which are allowed to depend on (θ_0,ϕ_0)),

$$\int \left|P(T_n<\theta_n;\ \theta_n,\phi) - \tfrac{1}{2}\right| dH_{\theta_n}(\phi) \to 0$$

as $n=m_r\to\infty$, provided θ_0 is not in N. It is immediate that

$$\left|\int P(T_n<\theta_n;\ \theta_n,H_{\theta_n}) - \tfrac{1}{2}\right| \to 0$$

along the subsequence m_r, except for a null set which is independent of the distributions H_{θ_n} which were chosen. It follows that the condition (12) of Theorem 5.7.1 holds

for θ_0 not in N and all functionals H which generate scores which satisfy the convergence conditions. Hence for all (θ_0,ϕ_0), with θ_0 not in N, if T_n is CAN at (θ_0,ϕ_0), then its asymptotic variance is no smaller than

$$\inf_H I_H^{-1}(\theta_0;\phi_0) = I^{*-1}(\theta_0;\phi_0).$$

5.9 The asymptotic information I*. It is conjectured that in a smooth parametric family the I* information for the Type I model is the same as the limit as $n\to\infty$ of the K-information. It is true for Models A, D, and E, where they are both equal to I_p, the partial information. The reason that they are known to equal those models is as follows: First, in each the partial likelihood is a directional score, so that $I^* \leq I_p$. Secondly, it will be shown below that $K \leq I^*$. Since in these models $K=I_p$, we must have $K=I^*=I_p$.

Suppose that the continuity assumption of lemma 4.4.2 is satisfied and Φ is sigma-compact, so that the minimization involved in K* can take place over P, the family of probability measures on Φ. For upper score generating functional H_θ, assume that $(L(\theta,H_\theta) - 1)^2/(\theta-\theta_0)^2$ is uniformly integrable with respect to the null distribution. Then

$$\frac{K^*(\theta)}{(\theta-\theta_0)^2} = \frac{\text{Var } L(\theta,H_\theta)}{(\theta-\theta_0)^2} \to E(U_H)^2 = I_H.$$

It follows that $K^{+} \leq I_H$ for all upper H. Similarly $K^{-} \leq I_H$ for all lower H.

On the other hand, suppose that for each $\theta > \theta_0$ there exists a least favorable distribution G_θ in the sense of definition 4.5.4. Then if $L(\theta, G_\theta)$ generates a directional score at (θ_0, ϕ_0), we have $K^{+} = I_G$. The conjecture is based on the presumption that if the density f is smooth enough, then G_θ and the likelihood ratio it generates will be also.

In the Type II model, one can make a similar argument, except that in this case the I* information equals the J-information provided that for the least favorable distribution Q_θ in P, $L(\theta, Q_\theta) = f(X;\theta, Q_\theta)/f(X;\theta_0, Q_0)$ is smooth enough to qualify as a directional-score-generating likelihood ratio.

5.10 Some questions. Thus far the theory has moved smoothly along the path laid by Bahadur. But the difficulties have merely been delayed, as will be seen shortly.

5.10.1 Do the proposed estimators meet the bound? (The key models) The answer is yes in the Type I Models A, B, D, and E, as the partial maximum likelihood estimates involved have variance equal to the inverse of the partial Fisher information. Recall that this is generally true

for conditional maximum likelihood estimates (see 3.6.4) and marginal likelihood estimates (because of their i.i.d. nature).

The answer is still yes in all the Type II models, where for Model C the bound is restricted to null points (θ_0, Q_0) such that Q_0 has infinitely many points of increase.

5.10.2 Are the proposed estimators CAN-UMU? (Type I model) In Models A, D, and E, since the estimators s^2_{nk}, $\bar{Z}_{nk}$, and $\bar{W}_n$ are i.i.d. averages, they are CAN. They are UMU because their distributions do not depend on ϕ.

In Model B, if $\phi \to \pm\infty$, then $P(X+Y=1;\ \theta, \phi) \to 0$. It follows that if the sequence $\phi_1, \phi_2, \phi_3, \ldots$ trends large in absolute value at a fast enough rate, then

$$\sum_{i=1}^{\infty} P(X_i+Y_i = 1;\ \theta, \phi_i) < \infty.$$

It follows by the Borel-Cantelli lemma (e.g., Rao {1973}, p. 137), that

$$P(X_i+Y_i = 1 \text{ infinitely often}) = 0.$$

Since the conditional distribution is binomial, with effective sample size equal to the number of times $X_i+Y_i=1$,

it follows that the CMLE T_n is not consistent. The inconsistency is itself not a problem as the lower bound theorem 5.8.2 did not require the estimate to be universally CAN. However, it was required to be universally UMU. Do we expect the CMLE T_n of Model B to be median unbiased for finite effective sample sizes? The answer is no. And is this a reasonable requirement for a family of estimates? It is not, as for some sequences of nuisance parameters there well may be no consistent estimators. In Model B the information about θ is effectively washing out as $\phi \to \infty$.

The discussion about how the UMU requirement might be weakened is postponed until the analogous problem in the Type II model is discussed.

5.10.3 Are the Type II estimators CAN-UMU? The remarks of 5.10.2 still hold for Models A, D, and E: yes, they are.

In Model B the requirement that

$$\sup_{Q} \left(\left| P(T_n < \theta \; ; \; \theta, Q) - \tfrac{1}{2} \right| \right) \to 0 \text{ as } n \to \infty$$

appears to be too demanding. The reason is that by using Q with high tail probabilities,

$$P \{X_i + Y_i = 1; \; \theta, Q\}$$

can be made arbitrarily small. The smaller this probability is, the slower the approach to median-unbiasedness of any estimator based on the conditional likelihood would be, since the effective sample size of the binomial distribution will grow more slowly.

5.11 UMU on compact sets. Andersen's solution to the problem of ill-trending sequences is to restrict the ϕ-sequence so that each ϕ_i comes from a compact set Φ_0. This assured a uniform approach to normality of the CMLE, T_n, and so, in particular,

$$(15) \qquad \sup \{ \, |P(T_n < \theta;\ \theta, \phi^{(\infty)}) - \tfrac{1}{2}| \, \} \to 0 \text{ as } n \to \infty.$$

For the CMLE, (15) is true for every compact set Φ_0 under the conditions of Theorem 3.7.2.

5.11.1 Definition. If T_n is an estimator which satisfies (15) for every compact set Φ_0 contained in Φ, then T_n is uniformly median unbiased over compact sets (UMUC).

There are two approaches which might now be taken. In the first, we restrict attention to the compact Φ_0. The preceding theory applies in this case, with the exception that in deriving the directional scores for a null

$(\theta_0, \phi_0^{(\infty)})$, $\phi_0^{(\infty)}$ must be in Φ_0^∞ and the mixing distributions used to create the directional score must have support in Φ_0^∞. The information $I^*(\Phi_0) := \inf \{ I_H(\theta_0; \phi_0^{(\infty)}) : H_\theta$ with support in $\Phi_0 \}$ then determines a lower bound for estimators which satisfy (15), the UMU condition for the set Φ_0. Suppose that for a sequence of compact sets Φ_k increasing to Φ we have

$$I^*(\Phi_k) \to I^*(\Phi) \quad \text{as } k \to \infty.$$

Then the following intuition is reinforced: Even if there exists an *a priori* compact bound on the ϕ-sequence, if that compact set is sufficiently large, the resulting lower bound is indistinguishable from the one for unbounded sequences. Hence, in particular, if that bound is unknown, then $I^*(\Phi)$ is the best bound one can use.

Unfortunately, this approach cannot be taken in this chapter. The regularity and convergence conditions of 5.4 and 5.5 appear to make the theory very difficult. This approach will, however, be used in Chapter 6, where asymptotic effective variance is used as an optimality criteria.

The route which will be taken here is to find sufficient conditions under which UMUC implies the lower bound theorem 5.8.2. Recall that a sufficient condition for the lower bound to hold at $(\theta_0, \phi_0^{(\infty)})$ was that for the score

distributions H_θ to be used

$$(16) \qquad \liminf_{n \to \infty} P(T_n < \theta_n \; ; \; \theta_n, H_{\theta_n}) \leq \tfrac{1}{2}.$$

We wish to know when T_n being UMUC will imply that this holds for almost all θ_0, for the set of H_θ we desire to use.

Fortunately, in two of our models the criterion (16) is no obstacle. First, in Model E, the space Φ consists of two points and so is compact. In Model B, the mixing distributions are themselves of common compact support provided the Φ-sequence is bounded. The proof of this is now sketched.

Recall (Section 4.9.6) that the nuisance eliminating distributions H_θ for the remainder likelihood (which was the marginal of X+Y) put mass on two points ϕ^+ and ϕ^- such that

$$r(1; \theta, \phi^-) = r(1; \theta_0, \phi_0) = r(1; \theta, \phi^+).$$

If $\phi^{(\infty)}$ is in Φ_0^∞ (and hence bounded), then $\{r(1;\theta,\phi_i)\}$ is bounded away from zero, and hence so are $\{r(1;\theta,\phi_i^-\}$ and $\{r(1:\theta,\phi_i^+\}$. But because the latter two sequences are bounded away from 0, the ϕ-sequences within them must be bounded away from $\pm\infty$. That is, the nuisance eliminating

distributions have support in compact set Φ_1. Applying the results of 5.9 to the parameter space $\Theta \times \Phi_1^\infty$ gives the result (16) for almost all θ_0 and all $\phi_0^{(\infty)}$ in Φ_0^∞.

Suppose now that H_θ does not necessarily have compact support. Because of the UMUC condition, if we define

$$h_{nk}(\theta) := \sup_{\Phi_k^\infty} \{|P(T_n < \theta;\ \theta, \phi^{(\infty)}) - \tfrac{1}{2}|\} \text{if } \theta \text{ in } \Theta$$
$$:= 0 \text{ else,}$$

where Φ_k, $k = 1,2,3,\ldots$ is a sequence of compact sets increasing to sigma-comapct Φ, then by lemma 5.8.1 for each k there exists a null set N_k and a sequence of integers m_r which depends on k such that $\lim h_{m_r k}(\theta_n) = 0$ if $\theta \notin N_k$. From the proof of Theorem 5.8.2, it follows that for any probability measure G_n with support in Φ_k

$$(17) \qquad |P(T_n < \theta_n\ ;\ \theta_n, G_n) - \tfrac{1}{2}| \to 0 \text{ as } n = m_r \to \infty$$

for all θ_0 not in N_k.

For any measure H_θ on Φ^n, let $H_{\theta k} = H_\theta\ (\ .\ |\ \Phi_k^n)$, the conditional probability measure. Because $H_{\theta k}$ has support in Φ_k^n, application of (17) yields

$$(18) \qquad \liminf_{n\to\infty} \ P(T_n < \theta_n;\ \theta_n, H_{\theta_n k}) \leq \tfrac{1}{2},\ \theta_0 \notin N_k.$$

Does this last inequality imply (16)? The following lemma provides a crude measure of the difference between the probabilities to be compared:

$$(19) \qquad a_{nk} := |P(T_n\ \theta_n;\theta_n,H_{\theta_n k}) - P(T_n\ \theta_n;\theta_n,H_{\theta_n})|$$

5.11.2 Lemma. For any measurable set B, if H_0 is the conditional probability measure $H(.|A)$, then

$$|P(B;\ \theta,H) - P(B;\ \theta,H_0)| \leq 2\ H(A^C)/H(A).$$

Proof. Let $H(A) = \delta$. Let I be the indicator function for the set B. First notice that

$$f(x;\theta,H_0) = \delta^{-1} \int_A f(x;\theta,\phi)\ dH$$

$$= \delta^{-1}(f(x;\theta,H) - \int_{A^C} f(x;\theta,\phi)\ dH).$$

Thus we have

$$|P(B;\theta,H) - P(B;\theta,H_0)| =$$

$$|\delta^{-1} \int I\ \{\delta f(x;\theta,H) - f(x;\theta,H) + \int_{A^C} f(x;\theta,\phi)\}d\nu|$$

$$\leq \delta^{-1}\int I(1-\delta)f(x;\theta,H)d\nu + \delta^{-1}\int\int_{A^C} f(x;\theta,\phi)dH\ d\nu,$$

by the triangle inequality. Substituting 1 for I, then applying Fubini's Theorem yields

$$= \delta^{-1}(1-\delta) \int f(x;\theta,H)\, d\nu + \delta^{-1} \int\int_{A^c} f(x;\theta,\phi)\, dH d\nu$$

$$= 2\,(1-\delta)/\delta.$$

This is what was to be shown.

A straightforward application of Lemma 5.11.2 is the following

$$a_{nk} \leq 2(H_{\theta_n}^{-1}(\phi_k^n) - 1), \tag{19}$$

where a_{nk} is defined in (18).

5.11.3 Theorem. If

$$\lim_{k\to\infty} \liminf_{n\to\infty} H_{\theta_n}(\phi_k^n) = 1 \tag{20}$$

and if T_n is UMUC, then I_H^{-1} is a lower bound for the asymptotic variance of T_n at all parameter points at which T_n is CAN, provided $\phi_0^{(\infty)} \varepsilon \cup \Phi_k^\infty$ and $\theta_0 \notin N$, a set of Lebesgue measure zero.

Proof. It will be shown that the inequality (12) holds at the stated parameter points, where $N=\cup N_k$. Suppose $\theta_0 \notin N$. Assumption (20), together with Equation (19) yields

$$\lim_{k\to\infty} \limsup_{n\to\infty} a_{nk} = 0$$

Let ε be positive. Choose and fix k so large that $\limsup_{n\to\infty} a_{nk}<\varepsilon$. Let n_0 be given. There exists n_1, without loss of generality greater than n_0, such that for all $n>n_1$, $a_{nk}<2\varepsilon$. Also, from (18), there exists $n_2>n_0$ such that

$$P(T_n<\theta_n \; ; \; \theta_n, H_{\theta_n k}) \leq \tfrac{1}{2} + \varepsilon.$$

It follows that for $n = n_2 > n_0$

$$P(T_n<\theta_n; \; \theta_n, H_{\theta_n}) \leq P(T_n<\theta_n; \; \theta_n, H_{\theta_n k}) + a_{nk}$$

$$\leq \tfrac{1}{2} + 3\varepsilon.$$

Since ε was arbitrary, Equation (12) holds.

5.11.4 Application. Suppose that in Model A, the one way ANOVA, the nuisance parameter is $\phi_0^{(\infty)}$, with entries ϕ_{i0}

bounded by c in absolute value. Let A_k be the closed interval from -k to k. The nuisance eliminating distribution for the marginal likelihood of the cell means is $H_{\theta_n} = \prod_{i=1}^{n} N(\phi_{i0}, n^{-\frac{1}{2}})$. Let Z be a standard normal random variable. The generalized Chebyshev inequality (e.g., Chung {1974}, p. 48) yields a bound on the tail probabilities of the distribution H_{θ_n}:

$$p_n := P(n^{-\frac{1}{4}}|Z| \geq r) \leq EZ^8 \, n^{-2} r^{-8}.$$

The interval A_{r+c} contains a closed interval of width r centered above each mean ϕ_{i0}, as all the means are contained in A_c. Denote the constant $r^{-8}EZ^8$ by a. Then

$$H_{\theta_n}(A^n_{r+c}) = \sum_{i=1}^{n} \{1-P(|n^{-\frac{1}{4}} Z + \theta_{i0}| \geq r+c)\}$$

$$\geq (1-p_n)^n$$

$$\geq (1-\frac{a/n}{n})^n.$$

It is well known that if $b_n \to b$, then $(1-b_n/n)^n \to e^{-b}$. It follows that $H_{\theta_n}(A^n_{r+c}) \to 1$ as $n \to \infty$. It follows that Theorem 5.11.3 holds. Notice that in this case all that was needed to achieve (12) was UMU on a single compact set A_{r+c} which contained the null sequence, with r arbitrarily small. The

implication is that given an *a priori* bound on the means, say c_o, for all mean sequences which have a bound strictly less than c_o, the estimate s^2_{nk} is optimal in the sense that I_p^{-1} is a lower bound for such sequences for almost all θ.

5.11.5 Application. In the preceding application the criterion (20) was verified using only the moments of H. Thus this technique is a natural accessory to the approach of Section 4.9.7, where the existence of a nuisance eliminating distribution H was demonstrated by the computation of its moments.

Unfortunately, in Model D, the tails of the H_{θ_n} are not shrinking fast enough. As can be seen from equation (73) of Section 4.9.8, the k-th moment goes to zero at the rate $|\theta_n - \theta| = n^{-\frac{1}{2}}$. (In the normal example, it went down at $n^{-k/4}$.) The use of the Chebyshev inequality as in the previous example fails, and it is conjectured that in this case (20) fails. Thus without further investigation, the UMU requirement cannot be reduced to UMUC and still retain the partial likelihood lower bound.

5.12 UMU in the Type II model. For the Type II model the natural analogy to the UMUC requirement for the Type I model is

(21) $$\sup_{Q\in P^*} \{|P(T_n<\theta\ ;\ \theta,Q) - \tfrac{1}{2}|\} \to 0 \text{ as } n \to \infty\ ,$$

where P^* consists of those measures on Φ with compact support. Estimators which satisfy (21) will also be called UMUC, with the context used to distinguish between Type I UMUC and Type II UMUC.

For a given null Q_0 with compact support, recall the nuisance eliminating distribution

$$dQ := dH_\theta^\phi\ dQ_0(\phi)$$

from Lemma 4.8.5, where H_θ^ϕ is a Type I nuisance eliminating distribution. The measure Q is not of compact support unless H_θ^ϕ is. In Models B and E, since the nuisance eliminating distributions are of compact support, UMUC is sufficient to yield lower bounds based on the partial informations. The arguments are similar to those of Section 5.11.

If the H_θ^ϕ are not of compact support, use compact support Φ_k for the conditional measure $H_{\theta k}^\phi := H_\theta^\phi(\ |\Phi_k)$. Let $dQ_{nk} = dH_{\theta_n k}^\phi\ dQ_0(\phi)$. Judicious use of Lemma 5.11.2 then yields

$$|P(\ T_n<\theta_n;\ \theta_n,Q_n) - P(T_n<\theta_n;\ \theta_n,Q_{nk})|$$

$$= \left| \int P(T_n < \theta_n ; \theta_n, H^{\phi}_{\theta_n}{}^{(n)}) - P(T_n < \theta_n ; \theta_n, H^{\phi}_{\theta_n k}{}^{(n)}) dQ_0^n(\phi^{(n)}) \right|$$

$$= \left| \int (H^{\phi i}_{\theta_n}(\Phi_k))^{-1} dQ_0(\phi_1) \ldots dQ_0(\phi_n) - 1 \right.$$

$$= \left(\int H^{\phi}_{\theta_n}(\Phi_k)\, dQ(\phi) \right)^{-n} - 1$$

The following analogy to Theorem 5.11.3 then holds:

5.12.1 Theorem. (Type II model) If for all $Q_0 \in P^*$

$$\lim_k \lim_n \inf \left(\int H^{\phi}_{\theta_n}(\Phi_k)\, dQ_0 \right)^n = 1,$$

where H^{ϕ}_{θ} are nuisance eliminating for the factorization $f = pr$, and the partial scores satisfy the regularity conditions of 5.4.1, then the asymptotic variance of a UMUC estimate T_n is not smaller than $I_p^{-1}(\theta_0; Q_0)$ except possibly for a Lebesgue null set of θ_0.

5.12.2 Application. In Model A, if Q has support in the closed interval from -c to +c, then let A_{r+c} be the closed interval from -r-c to +r+c. The arguments of 5.11.4 demonstrate that

$$\left(\int H^{\phi}_{\theta_n}(A_{r+c})\, dQ_0 \right)^n = \left(\int 1 - p_n\, dQ_0 \right)^n$$

$$= (1-p_n)^n \qquad \to 1 \text{ as } n \to \infty.$$

5.12.3 Other applications. As before, in Model D, the tails of the nuisance eliminating distributions fail to shrink fast enough as $n \to \infty$. In Model B, the measures are compact. In Model C, not enough is known about the nuisance eliminating distributions. (See 4.9.9.) In Model E, the space Φ is compact.

5.13 Chapter summary. The lower bound conclusions for the key models may be thus summarized:

5.13.1 Model A. The CMLE s^2_{nk} is optimal among all estimates which are CAN-UMUC, for both Type I and Type II models.

5.13.2 Model B. The CMLE achieves the lower bound for CAN-UMUC estimates, both Type I and Type II. In the Type I models, the bound exists only along those sequences for which $F_n \overset{w}{\to} F$.

5.13.3 Model C. There is a Type II lower bound for CAN-UMU estimates based on the partial likelihood. The reduction to CAN-UMUC estimates was not made, and since the CMLE is conjectured not to be UMU, its optimality is not established.

5.13.4 Model D. The CMLE achieves a lower bound for CAN-UMU estimates and is itself CAN-UMU. It is unknown if the same bound holds for CAN-UMUC estimates.

5.13.5 Model E. The invariant MLE achieves the lower bound for CAN-UMUC estimates.

CHAPTER VI

Lower Bounds For Asymptotic Effective Variance

6.1 Chapter Introduction. Like Chapter 5, the aim of this chapter is developing criteria with which to judge the asymptotic performance of estimators. It is based on the notion of asymptotic effective variance (see Section 1.6). The advantages and disadvantages of asymptotic effective variance (aev) as a measure of asymptotic efficiency are discussed in Section 6.2. The overall conclusion is that it is worthy of inclusion in this paper.

Section 6.3 ties together the Kullback-Leibler information distance (1.6.1) with the partial likelihood factorization. Then, an asymptotic model for which the Kullback-Leibler information is appropriate is presented in 6.4. That information distance is then related to the asymptotic theory of testing by a key lemma.

The lower bound theorem is proved in Section 6.5. The proof is based on a comparison of the Neyman-Pearson test (simple versus simple) with a test based on the estimator T_n, using the lemma of Section 6.4. The comparison yields the lower bound.

The next section, 6.6, discusses the computation of the lower bound in the case where there is a partial-simple nuisance factorization.

Sections 6.7 and 6.8 return to the problems discussed in 5.10. In some cases, the Type I space $\Theta \times \Phi^{\infty}$ is too large

for us to expect (or desire) estimators to be everywhere consistent. A reduction to compact sets is made in these sections. Finally, Section 6.9 revisits the key models to discuss the implications of the theoretical work.

6.2 Asymptotic Effective Variance. The Definitions 1.6.7 and 1.6.9 have the following interpretation in the nuisance parameter situation: Let Z be a standard normal random variable. Then $\tau_n = \tau_n(\varepsilon;\theta_0,\phi_0)$, the effective standard deviation of an estimator T_n for the function $g(\theta)$, is defined by

$$P(|T_n - g(\theta_0)| \geq \varepsilon) = P(|Z| \geq \varepsilon/\tau_n).$$

To compare two normal distributions with zero mean, it suffices to compare the variances. A comparison of the effective standard deviations $\tau_n(\varepsilon)$ of two estimators can yield a different outcome for each value of ε. In order to obtain a single limiting quantity to compare, namely the asymptotic effective variance, the limit infinum as ε goes to zero is used. If T_n is a sequence of estimators of $g(\theta)$, then the asymptotic effective variance at the null (θ_0,ϕ_0) of that sequence is

$$aev(T_n) := \liminf_{\varepsilon \to 0} \liminf_{n \to \infty} n\tau_n^2(\varepsilon).$$

Recall from Section 1.6 that Bahadur (1971) developed a lower bound theory for aev in i.i.d. models. Let us consider the advantages and disadvantages of using aev as an efficiency criterion.

The definition of asymptotic effective variance applies to any estimating sequence T_n, and in some ways this generality is a liability. When comparing the asymptotic variance of two CAN estimators, the intuitive feeling might be that for a large enough sample size both estimators are close to normality in distribution. Hence a comparison of variances would yield the more desirable estimate (with the caveat that the speed of the approach to normality is relevant). What does aev tell us? It is directly related to the quantity

$$\liminf_{n\to\infty} nP\{|T_n - g(\theta_0)| \geq \varepsilon\}.$$

(See 6.5.2.) This quantity indicates something about the probability that the standardized variable $n^{\frac{1}{2}}(T_n - g(\theta_0))$ has a large deviation ($n^{\frac{1}{2}}\varepsilon \to \infty$ as $n \to \infty$) from zero. It is doubtful that the probability of a large deviation is as important to the practicing statistician as the variance of his estimate. From a decision theoretic viewpoint, asymptotic variance is similar to mean square error risk, while aev measures risk for a loss which is 1 for large deviations

and 0 else.

Further, the aev concept is in some ways unattractive mathematically. The computation of the aev of an estimator can be difficult compared with computing its asymptotic variance. Also, the question of when the maximum likelihood estimate meets the lower bound (1.6.8) in the i.i.d. situation is only partially solved (Bahadur {1971}). Since the computation and theory of the aev are difficult for the MLE, one suspects that they will be more so for CMLE's or other partial MLE's.

There are, however, ways in which the concept of aev proves an advantage. First, since the concept applies to all estimators, the lower bound derived will hold for all consistent estimators. Next, there is no problem of superefficiency. Lower bound results will hold for all null points simultaneously, without any "almost all θ_0" restriction. This result is obtained without uniformity requirements on the estimators. Also, the theory corresponds to the asymptotic theory of testing of Bahadur. It will be seen that aev tells us something about the properties of the estimator when used as a test statistic.

The final advantage of the aev concept is that theorems may be proved without regularity and convergence

assumptions such as those of Sections 5.4 and 5.5. As a result, the problem of ill-mannered sequences which was dealt with in Sections 5.10, 5.11, and 5.12 of the preceding chapter can be treated here in a more satisfactory manner.

6.3 Partial Likelihoods. In the nuisance parameter setting, the Kullback-Leibler information distance (1.6.1) between (θ,ϕ) and (θ_0,ϕ_0) will be expressed as

$$\kappa(\theta,\phi) := E(L(\theta,\phi) \ln L(\theta,\phi)) \text{ for } (\theta,\phi)<<(\theta_0,\phi_0),$$

$$:= \infty \text{ else,}$$

where $L(\theta,\phi) = f(X;\theta,\phi)/f(X;\theta_0,\phi_0)$. The greek letter κ will be used here to avoid confusion with the K-information of Chapter 4. If the parameter space is extended to $\theta\times P$ by mixing distributions on Φ, the κ-distance between (θ,H) and (θ_0,ϕ_0) is $\kappa(\theta,H) := E(L(\theta,H) \ln L(\theta,H))$.

Consider now the impact of using as mixing distributions the nuisance eliminating distributions H for a partial likelihood factorization f = pr. In that case:

$$\kappa(\theta,H_\theta) = E\{L_p(\theta) \ln L_p(\theta)\} := \kappa_p(\theta).$$

The following lemma provides a connection between the

partial Fisher's information I_p and the partial Kullback-Leibler information distance $\kappa_p(\theta)$. It is analogous to Theorem 1.6.3, which related the κ-distance to the full Fisher information.

6.3.1 Lemma. Suppose p is a partial likelihood such that $E\, L_p' = 0$, $E\, L_p'' = 0$, and the function $\kappa_p(\theta)$ is twice differentiable in θ at θ_0, with

$$\partial\kappa_p(\theta)/\partial\theta = E\partial L_p \ln L_p/\partial\theta$$

and

$$\partial^2\kappa_p(\theta)/\partial\theta = E\partial^2 L_p \ln L_p/\partial\theta^2.$$

Then

$$\lim_{\theta\to\theta_0} \kappa_p(\theta)/\tfrac{1}{2}(\theta-\theta_0)^2 = I_p(\theta_0).$$

Proof. Apply L'Hopital's rule. The first derivative of $\kappa_p(\theta)$ at θ_0 is zero, and the second is $I_p(\theta_0)$:

$$\kappa_p'(\theta) = E(L_p' \ln L_p + L_p')$$

$$\kappa_p''(\theta) = E(L_p'' \ln L_p + (L_p')^2 + L_p'').$$

Since $L_p(\theta_0) = 1$, the result is straightforward.

The following lemma suggests that the above result holds quite generally if p is from an exponential family.

6.3.2 Lemma. If $p(x;\theta) = c(\lambda) \exp(\lambda^t S)$, where $S = S(x)$ and $\lambda=\lambda(\theta)$ are vectors, then

$$E\, L_p \ln L_p = \frac{c(\lambda)}{c(\lambda_0)} \Sigma\, (\lambda_i(\theta) - \lambda_i(\theta_0)\partial c^{-1}(\lambda)/\partial\lambda;$$

where $\lambda_0 = \lambda(\theta_0)$.

Proof. The following string of equalities uses two key results: the theorem of Lehmann ({1959}, p. 52) concerning the differentiability under an integral of an exponential function, and the fact that $E\, L_p = 1$ implies that

$$E \exp(\lambda^t S - \lambda_0^t S) = c(\lambda_0)/c(\lambda).$$

The equalities:

$$EL_p \ln L_p = c(\lambda)c^{-1}(\lambda_0)E\{\Sigma(\lambda_i-\lambda_{i0}) \frac{\partial}{\partial\lambda_i} \exp S^t(\lambda-\lambda_0)\}$$

$$= c(\lambda)c^{-1}(\lambda_0)\Sigma(\lambda_i-\lambda_{i0}) \frac{\partial}{\partial\lambda_i} E\{\exp S^t(\lambda-\lambda_0)\}$$

$$= c(\lambda)c^{-1}(\lambda_0)\Sigma(\lambda_i-\lambda_{i0}) \frac{\partial}{\partial\lambda_i} c^{-1}(\lambda).$$

6.4 Asymptotic Model. As in Chapter 5, the general asymptotic model will be for $X^{(n)}$ to have marginal density

$$f^{(n)}(X^{(n)};\ \theta,\phi) = \prod_{i=1}^{n} f_i(X^{(i)};\ \theta,\phi)$$

with respect to measure $\nu^{(n)}$. Also, as before, restriction must be placed on the use of mixing distributions so that the log likelihood ratio $\ln L^{(n)}$ is a sum. In this case it is not needed for a Central Limit Theorem, but rather for a Weak Law of Large Numbers. Hence in the Type I model the mixing distributions $H_\theta^{(n)}$ will be product measures, $\prod_{i=1}^{n} H_\theta^{\phi_i}$, where the superscript ϕ_i indicates that the mixing distribution corresponds to null (θ_0,ϕ_i) for observation X_i. In this case then

$$\ln L^{(n)}(x^{(n)};\ \theta,H_\theta^{(n)}) = \Sigma \ln L(X_i;\ \theta,H_\theta^{\phi_i}).$$

In the Type II model, or any other where each density f_i depends on the entire muisance ϕ, one may use no mixing, only least favorable parameters ϕ_θ. (In the Type II model, of course, the mixing distributions are themselves the parameters.)

6.4.1 Assumptions. Thus for the general model it is assumed that the log likelihood ratio is a sum for some

alternative (θ,H) to the null (θ_0,ϕ_0):

$$\ln L^{(n)}(\theta,H) = \Sigma \ln L_i(\theta,H).$$

It is further assumed that

(a) $\kappa_i = E(\ln L_i ; \theta,H) < \infty$, and
(b) $n^{-1}(\Sigma \ln L_i - \kappa_i) \xrightarrow{P} 0$ under (θ,H).

The Assumption 6.4.1.b is a WLLN result. Thus if one is using for H a nuisance eliminating distribution, in a Type II model or in a Type I model where P is a marginal likelihood, $\ln L_i = \ln L_p(X_i)$, $i = 1,2,3,\ldots$ is an i.i.d. sequence, so the assumption holds provided the mean $\kappa_p(\theta)$ is finite. Other Type I cases will be discussed later when the subject of null sequences is covered (Section 6.7).

The above assumptions yield the following lemma, a generalization of Lemma 1.6.6 (Bahadur).

6.4.2 Lemma. Let $\alpha_n(\beta)$ be the size of the least size test of power β for testing the null hypothesis $P_0 := P_{(\theta_0,\phi_0)}$ against the alternative $P_1 := P_{(\theta,H)}$ based on $X^{(n)}$. For all

$\beta\varepsilon(0,1)$

$$\liminf_{n\to\infty} n^{-1} \ln \alpha_n(\beta) \geq - \limsup_{n\to\infty} n^{-1}\Sigma\kappa_i .$$

Proof. (Based on Bahadur {1971}) Let ω_n be the least size test (possibly randomized). Let d_n be any sequence of positive constants, and let $J_n = I\{L^{(n)} \leq d_n\}$. Then

(1)
$$\begin{aligned} \alpha_n(\beta) = E(\omega_n) &= \int \omega_n \, dP_0 \\ &= \int J_n \omega_n \, dP_0 \\ &= d_n^{-1} \int J_n \omega_n L^{(n)} \, dP_0 \\ &= d_n^{-1} \int J_n \omega_n \, dP_1 \\ &= \{(1-\beta) - P_1(L^{(n)} > d_n)\} d_n^{-1}. \end{aligned}$$

For the sequence d_n we will use $\exp(\Sigma\kappa_i + n\varepsilon)$, for $\varepsilon > 0$.

Notice then that

$$\begin{aligned} n^{-1}(\ln L^{(n)} - \ln d_n) &= n^{-1}(\Sigma(\ln L_i - K_i) - n\varepsilon) \\ &\xrightarrow{P} -\varepsilon \text{ under } (\theta, Q), \end{aligned}$$

by Assumption 6.4.1.b. It follows that $P_1(L^{(n)}>d_n)$ is $o(1)$ as $n\to\infty$. The inequality starting at (1) then yields:

$$n^{-1} \ln \alpha_n(\beta) \geq n^{-1} \ln \{d_n^{-1}(1-\frac{o(1)}{1-\beta})(1-\beta)\}$$

$$= -n^{-1}\Sigma\kappa_i - \varepsilon + n^{-1} \ln(1-o(1)) + o(1).$$

Taking the limes inferior of both sides of the last inequality gives the desired result, since ε was arbitrary.

6.4.3 Definition. Let $\bar{\kappa} := \limsup_{n\to\infty} n^{-1}\Sigma\kappa_i$ be the asymptotic κ-information.

6.5 The Lower Bound. We are now prepared to generalize Bahadur's lower bound theorem for asymptotic effective variance from the i.i.d. case to the general model. As in Chapter 5, the estimator T_n is turned into a test statistic. Let

$$(2) \qquad J_n(\varepsilon) = I(|T_n - g(\theta_0)| \geq \varepsilon).$$

Then, when viewed as a test, $J_n(\varepsilon)$ has size $E(J_n(\varepsilon)) := \alpha_n(\varepsilon)$. The lemmas below will enable us to relate $\alpha_n(\varepsilon)$ to the effective standard deviations and the aev (T_n).

6.5.1 Lemma. For t positive, Z a standard normal random variable,

$$(t^{-1}-t^{-3})(2t)^{-1}\exp(-t^2/2) < P(|Z|\geq t)$$

$$< t^{-1}(2\pi)^{-1}\exp(-t^2/2).$$

Proof. See Feller (1968), p. 175.

6.5.2 Lemma. Suppose $\liminf \alpha_n(\varepsilon)>0$ for some $\varepsilon>0$. If T_n is consistent for $g(\theta)$ in (θ_0,ϕ_0), then

$$\text{aev}\ (T_n) = -2\{\liminf_{\varepsilon\to 0} \liminf_{n\to\infty} \{(n\varepsilon^2)\ln \alpha_n(\varepsilon)\}\}^{-1}.$$

Proof. (Bahadur {1971}) Since $\alpha_n(\varepsilon)$ is strictly positive for n large, and since $\alpha_n(\varepsilon) = P\{|Z| \geq \varepsilon/\tau_n(\varepsilon)\}$ by definition, it follows that τ_n is strictly positive (possibly infinite). Since T_n is consistent, $\tau_n(\varepsilon)\to 0$ as $n\to\infty$. Since τ_n is positive, but goes to zero, Lemma 6.5.1 above yields

$$\ln \alpha_n(\varepsilon) = -\varepsilon^2/2\tau_n^2(\varepsilon)(1+o(1)) \text{ as } n\to\infty.$$

The lemma follows by an appropriate modification of the last equation.

The following easy lemma is the only step required to insert the mixing distribution H into the Bahadur theory. Let $\beta_n(\varepsilon;\theta,\phi) = E(J_n(\varepsilon);\ \theta,\phi)$ and $\beta_n(\varepsilon;\ \theta,H) := E_1(J_n(\varepsilon))$, where J_n is defined as in (2), and E_1 is expectation with respect to P_1.

6.5.3 Lemma. If T_n is consistent for $g(\theta)$ in (θ,ϕ), then $\beta_n(\varepsilon;\theta,H) \to 1$ for $|g(\theta)-g(\theta_0)|>\varepsilon>0$.

Proof. By consistency $\beta_n(\varepsilon;\ \theta,\phi)\to 1$ for all ϕ. Thus if H is any probability measure on Φ, the dominated convergence theorem implies:

$$\beta_n(\theta,H) = \int \beta_n(\varepsilon;\ \theta,\phi)\ dH \to 1 \text{ as } n\to\infty.$$

Next, we establish the information measure which will play the role that the K-information played in Chapter 4 and the I* information played in Chapter 5: the denominator in a Cramer-Rao-type lower bound.

6.5.4 Definition. Let H_θ be a function which assigns a a probability measure H_θ to all θ in a neighborhood adjoining θ_0. Define the direction-H κ*-information to be

$$\kappa_H^* = \lim_{\theta \to \theta_0^{+(-)}} \sup \ \overline{\kappa}(\theta, H_\theta)/\tfrac{1}{2}(\theta-\theta_0)^2 .$$

6.5.5 Theorem. For any H_θ satisfying 6.4.1 and any consistent estimator T_n

$$\text{(3)} \qquad \text{aev}(T_n) \geq (\, g'(\theta_0))^2/\kappa_H^* .$$

Proof. (Modification of Bahadur {1971}) For $p\varepsilon(0,1)$, let $\varepsilon := p|g(\theta)-g(\theta_0)|$. Let $J_n(\varepsilon)$ be the test statistic defined in (2). By Lemma 6.5.3, the power of this test, $\beta_n(\varepsilon;\theta,H_\theta)$, goes to 1 as $n\to\infty$. In particular, then, $\beta_n(\varepsilon;\theta,H_\theta)>\frac{1}{2}$ for n larger than some n_0. Let $\alpha_n(\varepsilon) := E(J_n(\varepsilon))$ be the size of this test. A comparison with the size α_n^* of the least size test of power $\frac{1}{2}$ gives

$$\alpha_n(\varepsilon) \geq \alpha_n^* .$$

Hence by Lemma 6.4.2

$$\text{(4)} \qquad \liminf_{n\to\infty} \ n^{-1} \ln \alpha_n(\varepsilon) \geq \liminf_{n\to\infty} n^{-1} \ln \alpha_n^* \geq -\overline{\kappa}(\theta,H).$$

Thus, using the definition of ε,

$$(5)\qquad \liminf_{\varepsilon\to 0} \liminf_{n\to\infty} (n\varepsilon^2)^{-1} \ln \alpha_n(\varepsilon) \geq \liminf_{\theta\to\theta_0^{+(-)}} \frac{-\frac{1}{2}\overline{\kappa}(\theta, H_0)}{\frac{1}{2}(\theta-\theta_0)^2} \frac{(\theta-\theta_0)^2}{p^2(g(\theta)-g(\theta_0))^2}.$$

From the right hand term of (5) we obtain $-\frac{1}{2}\kappa_H^*/p^2(g'(\theta_0))^2$. Since p was arbitrary in (0,1), p = 1 maintains the inequality.

Next, consider the left hand term of (5). If $\kappa_H^* = \infty$, then (3) holds trivially. If not, then (5) implies that $\liminf_{n\to\infty} \alpha_n(\varepsilon)>0$ for ε sufficiently small. Hence Lemma 6.5.2 applies to the left hand term of (5), yielding the proper inequality.

6.6 Computing κ_H^* for the simple nuisance problem. Suppose that the X_i's are independent and that their common density f has a partial likelihood factorization f = pr. Further, suppose that r is a simple nuisance. We now consider the computation of κ_H^* for the nuisance eliminating distribution. In the Type II model, if Q_θ is nuisance eliminating, then

$$\overline{\kappa}(\theta, Q_\theta) = \kappa_p(\theta)$$

and so an application of Lemma 6.3.1 yields $\kappa_Q^* = I_p$.

In the Type I model, does the κ^* information reduce to the partial Fisher's information? Let H_θ^ϕ be nuisance

eliminating for null (θ_0,ϕ), let $\kappa_p^{\phi}(\theta)$ be the partial κ-distance for null (θ_0,ϕ), and, in general, let the superscript ϕ denote that the null to be used is (θ_0,ϕ). Suppose that lemma 6.3.1 holds for each null (θ_0,ϕ_i) for null sequence $\phi^{(\infty)}=(\phi_1,\phi_2,\ldots)$. Does it follow that $\kappa_H^* = \limsup n^{-1} \Sigma I_p^{\phi_i}(\theta_0)$? It is true when the partials are marginal distributions as there is no dependence on index i. It is not necessarily true otherwise, as two limiting operations must otherwise be reversed. This will be seen in the following lemma.

Let $\overset{u}{\to}$ denote uniform convergence with respect to the variable ϕ.

6.6.1 Lemma (Type I model) If

$$\kappa_p^{\phi}(\theta)/\tfrac{1}{2}(\theta-\theta_0)^2 \overset{u}{\to} I_p^{\phi}(\theta_0) \quad \text{as } \theta\to\theta_0^{+(-)}$$

then

$$\kappa_H^* = \limsup n^{-1} \sum_{i=1}^{n} I_p^{\phi}(\theta) := \overline{I}_p.$$

Proof. By definition

$$\kappa_H^* = \limsup_{\theta \to \theta_0^+} \limsup_{n \to \infty} n^{-1} \Sigma \kappa_p^{\phi} i(\theta)/(\tfrac{1}{2})(\theta-\theta_0)^2 .$$

The given uniformity condition ensures that the two limes superior may be reversed in order.

The only Type I model of the non-marginal type with a simple nuisance is Model B. In that model

$$\kappa_p^{\phi}(\theta)/\tfrac{1}{2}(\theta-\theta_0)^2$$

$$= P\{X+Y=1;\ \theta_0,\phi\}\ E(L_p \ln L_p | X+Y=1)/\tfrac{1}{2}(\theta-\theta_0)^2 .$$

The term $P(X+Y=1;\ \theta_0,\phi)$ is constant in θ. The term

$$E(L_p \ln L_p | X+Y=1)/\tfrac{1}{2}(\theta-\theta_0)^2$$

does not depend on ϕ. Hence it converges uniformly in ϕ to the conditional Fisher's information $e^{\theta}/(1+e^{\theta})^2$. Lemma 6.6.2 is thus applicable for every ϕ-sequence.

6.7 The reduction to compact sets. The lower bound theorem 6.5.5 required that the estimator T_n be consistent for all (θ,ϕ), where in the Type I model ϕ means infinite sequences $\phi^{(\infty)}$ and in the Type II model ϕ means probability measures Q in P. As discussed in 5.10, it may be unreason-

able to expect consistency over such a large space. In fact, such a space is larger than we might consider reasonable for modelling many statistical problems.

Hence, as before, we will use for a Type I space $\Phi_B := \cup \Phi_k^{\infty}$, where compact Φ_k are increasing to Φ. If $\Phi=R$, then Φ_B consists of all bounded sequences. On this space the CMLE is, under regularity, consistent (see 3.5.4). In the type II model no reduction need be made for the sake of the CMLE, as it is consistent for θ in (θ,Q), for all Q (see 3.5.8). However, results about the space $\mathcal{P}_B$, the family of probability measures with compact support, will be a corollary to the considerations for Φ_B.

A byproduct of restricting to Φ_B is the satisfaction of Assumption 6.4.1.b. If the variance of $\ln L_i(\theta,H^{\phi})$ under (θ,H^{ϕ}) is continuous in ϕ, then it is bounded for ϕ from a compact set. It follows that the WLLN implies that assumption (e.g., by Chebyshev's Theorem, Rao {1973}, p. 112). In particular, this is true for Type I Model B.

The following theorem demonstrates an advantage of using aev theory over CAN theory in the nuisance parameter setting. It greatly simplifies the reduction to compact sets.

6.7.1 Theorem. If Φ_k is a sequence of compact sets increasing to Φ, and $H_k := H(\ .\ |\Phi_k)$ are the conditional

probability measures based on Φ_k, then

$$\lim_{k\to\infty} \kappa(\theta, H_k) = \kappa(\theta, H)$$

provided $\kappa(\theta, H)<\infty$.

Proof. Let $L_k := \int_{\Phi_k} L(\theta,\phi)\, dH$, $L = L(\theta,H)$. Notice that $L(\theta,H_k) = L_k/H(\Phi_k)$. Then

$$\kappa(\theta,H_k) = H^{-1}(\Phi_k)\, E(L_k \ln L_k) - \ln H(\Phi_k). \tag{6}$$

Since $H(\Phi_k)\to 1$ as $k\to\infty$, we need only show that:

$$E\, L_k \ln L_k \to K(\theta,H).$$

Let $J_k = I\{L_k \geq 1\}$. Then both L_k and J_k are monotonely increasing in k. Write

$$E\, L_k \ln L_k = E\, J_k L_k \ln L_k + E(1-J_k)\, L_k \ln L_k. \tag{7}$$

The function $J_k L_k \ln L_k$ is strictly increasing in k, with limit $(L \ln L)^+$, and positive, so the monotone convergence of the integral applies. The function $(1-J_k)\, L_k \ln L_k$ is bounded above as it is negative and bounded below by e^{-1},

because the function y ln y is so bounded. The dominated convergence theorem then applies to $(1-J_k)\ L_k \ln L_k \to (L \ln L)^-$.

Application of the convergence theorems to (7) yields the theorem.

6.7.2 Corollary. (Type II model) If Φ_k are compact sets increasing to Φ, then any estimator T_n consistent for $g(\theta)$ in $\Theta \times P_B$ satisfies

$$\text{aev}\ (T_n) \geq (g'(\theta_0))^2/\kappa^*_H$$

for all H which satisfy Assumptions 6.4.1.

Proof. Any estimator which is consistent on P_B is consistent for each H_k. Thus we may proceed in Theorem 6.5.5 until the inequality (4) using mixing distribution H_k. Here (4) becomes

$$\liminf\ n^{-1} \ln \alpha_n(\varepsilon) \geq -\kappa(\theta, H_k).$$

Applying the limiting result of 6.7.1 to this yields

$$\liminf\ n^{-1} \ln \alpha_n(\varepsilon) \geq -\kappa(\theta, H).$$

The rest of the proof of 6.5.5 proceeds as before.

The last corollary quite tidily disposes of the question of compact sets for the Type II model. Unfortunately, the Type I model turns out to be quite difficult at this point. The result needed for substitution in Equation (4) is

$$\overline{\kappa}(\theta, H_k^{\infty}) \to \overline{\kappa}(\theta, H^{\infty}).$$

Thus the following interchange is needed:

$$\lim_k \limsup_n n^{-1} \Sigma \kappa(\theta, H_k^{\phi} i) = \limsup_n \lim_k n^{-1} \Sigma \kappa(\theta, H^{\phi} i). \tag{8}$$

This will be true provided that the limiting result in Theorem 6.7.1 is uniform for ϕ in compact sets.

It should be pointed out that the difficulty here is largely mathematical, not conceptual. If the null parameter is $\phi^{(\infty)} = (\phi, \phi, \phi, \ldots)$, then the limit interchange in (8) is trivial. Similarly, if the sequence $(\phi_1, \phi_2, \phi_3, \ldots)$ contains only a finite (but arbitrarily large) number of different numbers, the limit is uniform over that set in Theorem 6.7.1, and (8) holds.

In the following, notation will be trimmed considerably because the variables θ and H will be fixed. In particular,

$$f_k^{\phi} := \int_{\Phi_k} f(X;\theta,\phi)\, dH_k, \quad f^{\phi} := f(X;\theta,H^{\phi}),$$
$$L_k^{\phi} := f_k^{\phi}/f(X;\theta_0,\phi), \quad L^{\phi} := f^{\phi}/f(X;\theta_0,\phi).$$

What is desired, then, is

$$\kappa^{\phi}(\theta,H_k^{\phi}) \overset{u}{\to} \kappa^{\phi}(\theta,H^{\phi}) \quad \text{as } k\to\infty. \tag{9}$$

The following uniform convergence results will be useful:

6.7.3 Uniform Convergence Criteria. (Rudin {1964}, pp. 136, 153)

(a) Suppose $f_n(\phi)$ are continuous functions on a compact space, and that f_n converge pointwise to f, another continuous function. If the f_n are monotonically decreasing at each ϕ as n increases, then $f_n \overset{u}{\to} f$.

(b) The sum of two uniformly convergent sequences is uniformly convergent.

(c) The product of two uniformly convergent

sequences of bounded functions is a uniformly convergent sequence.

The following assumptions are used for the Type I model. They will be simplified later for the simple nuisance model.

6.7.4 Continuity Assumptions. The following integrals exist and are continuous functions of the parameter ϕ for all k:

(a) $H^{\phi}(\Phi_k)$

(b) $\int f^{\phi}(\ln L^{\phi} - \ln L_k^{\phi})\, d\nu$

(c) $\int(f^{\phi} - f_k^{\phi})(\ln L^{\phi} - \ln L_k^{\phi})\, d\nu$

(d) $\int(f^{\phi} - f_k^{\phi})\, |\ln L^{\phi}|\, d\nu$.

6.7.5 Theorem. Under assumptions 6.7.4,

$$\kappa^{\phi}(\theta, H_k^{\phi}) \overset{u}{\to} \kappa^{\phi}(\theta, H^{\phi}) \quad \text{as } k \to \infty,$$

with the uniformity being over all compact sets.

Proof. A reexpression of (6) becomes

$$\kappa^{\phi}(\theta, H_{,k}^{\phi}) = (H^{\phi}(\Phi_k))^{-1} \int f_k^{\phi} \ln L_k^{\phi}\, d\nu - \ln H^{\phi}(\Phi_k).$$

Application of Assumption 6.7.4(a) with convergence criterion 6.7.3(a) gives

$$H^{\phi}(\phi_k) \overset{u}{\to} 1.$$

It follows, by 6.7.3(b) and (c), that it suffices to show that

$$\int f_k^{\phi} \ln L_k^{\phi} d\nu \overset{u}{\to} \int f^{\phi} \ln L^{\phi} \, d\nu = \kappa^{\phi}(\theta, H^{\phi}).$$

It will be shown that

$$(10) \qquad \int f^{\phi} \ln L^{\phi} \, d\nu - \int f_k^{\phi} \ln L_k^{\phi} \, d\nu$$

converges uniformly to zero by showing that it is the sum of three uniformly converging terms. Important to the argument is the monotonely increasing character of L_k^{ϕ} and f_k^{ϕ}. The difference (10) is the sum of three terms:

(a) $\int f^{\phi} (\ln L^{\phi} - \ln L_k^{\phi}) \, d\nu$

(b) $-\int (f^{\phi} - f_k^{\phi})(\ln L^{\phi} - \ln L_k^{\phi}) \, d\nu$

(c) $\int (f^{\phi} - f_k^{\phi}) \ln L^{\phi} \, d\nu$

The terms (a) and (b) are continuous in ϕ by assumption. Terms (a) and (b) are, respectively, monotonely decreasing

and the negative of a monotonely decreasing sequence. Hence they converge uniformly to zero. Term (c) is bounded by in absolute value

$$\int (f^{\phi} - f_k^{\phi}) \, |\ln L^{\phi}| \, dv$$

which by assumption is continuous, and which is monotonely decreasing to 0, hence converges uniformly. This implies (c) does also.

6.7.6 Corollary. Under Assumption 6.7.4,

$$\overline{\kappa}(\theta, H_k^{\infty}) \to \overline{\kappa}(\theta, H^{\infty})$$

for all null sequences in Φ_B.

6.7.7 Theorem. (Type I) If for $\Phi_k \to \Phi$, then $\kappa(\theta, H_{\theta_k}^{\infty}) \to \kappa(\theta, H_{\theta}^{\infty})$, then

$$\text{aev } (T_n) \geq (g'(\theta_0))^2 / \kappa_H^*$$

for all T_n consistent for $g(\theta)$ over $\Theta \times \Phi_B$.

Proof. As in the proof of 6.7.2.

The continuity assumptions thus provide a sufficient condition for the reduction to compact sets. They are, however, themselves awkward, and so a revised set will be found for partial likelihood-simple nuisance factorizations.

6.8 Continuity Assumptions and Simple Nuisance. In this section a simplified set of assumptions is found to replace the continuity assumptions 6.7.4 for the situation when the H_θ^ϕ are nuisance eliminating probability measures for a partial likelihood factorization f = pr. First, a lemma.

6.8.1 Lemma. If H_θ^ϕ are nuisance eliminating, if $f(x;\theta,\phi)$, f_k^ϕ, and $E(L_p^2 ; \theta_0,\phi)$ are continuous in ϕ, then continuity assumptions 6.7.4 (b), (c), and (d) are satisfied.

Proof. Since under the above hypothesis the integrands of the functions in the continuity assumptions are continuous, it will suffice to show that the limit as $\phi\to\phi_0$ and integral can be interchanged:

Suppose $f_k^\phi|\ln L_k^\phi|$ is bounded above by ν-integrable $M(\phi)$, such that $\int M(\phi)\, d\nu \to \int M(\phi_0)\, d\nu < \infty$ as $\phi\to\phi_0$. Since the same function is bounded below by zero, the extended dominated convergence theorem applies (e.g., Rao {1973}, p. 136). Hence $\int f_k^\phi|\ln L_k^\phi| \overset{d\nu}{\to} \int f_k^{\phi_0}|\ln L_k^{\phi_0}|d\nu$, as required

for continuity. Using $M(\phi)$ and $-M(\phi)$ as upper and lower dominating functions implies that $\int f_k^\phi \ln L_k^\phi \, d\nu \to \int f_{k_0}^\phi \ln L_{k_0}^\phi \, d\nu$.

Let $J_k := I\{L_k^\phi \geq 1\}$. Then

$$|f_k^\phi \ln L_k^\phi| = |J_k f_k^\phi \ln L_k^\phi + (1-J_k) f_k^\phi \ln L_k^\phi|$$

$$\leq |J_k f_k^\phi \ln L_k^\phi| + |(1-J_k) f(x;\theta_0,\phi) L_k^\phi \ln L_k^\phi|.$$

Now use the fact that for x greater than 1, $0 \leq \ln x \leq x$, and for x less than 1, $0 \geq x \ln x \geq -e^{-1}$, to get

$$\leq J_k f_k^\phi L_k^\phi + (1-J_k)\, e^{-1} f(X;\theta_0,\phi)$$

$$\leq f(\theta_0,\phi)(L^\phi)^2 + f(x;\theta_0,\phi).$$

If H_θ^ϕ is nuisance eliminating, then $L^\phi = L_p$, hence the bound

$$M(\phi) = f(x;\theta,\phi)\,(L_p^2 + 1).$$

By assumption $\int M(\phi) d\nu = EL_p^2 + 1$ is continuous in ϕ.

A similar bounding argument works for the other summands involved in continuity assumptions (b), (c), and (d).

6.8.2 Theorem. If the H_θ^ϕ are nuisance eliminating, then the following assumptions are sufficient for the lower bound

$$aev(T_n) \geq (g'(\theta_0))^2/\kappa_p^*,$$

where T_n is consistent for $g(\theta)$ in $\Theta\times\Phi_B$:

(a) $E(L_p^2;\ \theta_0,\phi)$ is continuous in ϕ.

(b) $H_\theta^\phi \xrightarrow{w} H_\theta^{\phi_0}$ as $\phi\to\phi_0$.

(c) Φ_k are continuity intervals for all H_θ^ϕ.

(d) $f(x;\theta,\phi)$ is continuous in ϕ.

Proof. It is demonstrated that continuity assumptions 6.7.4 hold. Assumption (a) follows from the equivalence of weak convergence and convergence on continuity intervals.

If $f(x;\theta,\phi)$ is continuous in ϕ, then for each (x,θ) it is bounded on Φ_k, and so by the definition of weak convergence,

$$\int_{\Phi_k} f(x;\theta,v)\ dH_\theta^\phi(v) \to \int_{\Phi_k} f(x;\ \theta,v)\ dH_\theta^{\phi_0}(v)$$

as $\phi\to\phi_0$. Hence f_k^ϕ is continuous. Application of Lemma

6.8.1 completes the proof.

6.9 Applications. There are two questions to be answered about each of the key models. First, does the partial likelihood generate a lower bound for aev in $\Theta\times\Phi_B$ or just $\Theta\times\Phi^\infty$? Secondly, does the partial MLE meet the bound? To the latter point the reader is reminded of Lemma 1.6.10: the sample mean of i.i.d. random variables, when used as an estimate of the mean, has for aev the variance, provided that the moment generating function exists in a neighborhood of zero.

6.9.1 Model A. This model meets the requirements of Theorem 6.8.2:

(a) Since the partial likelihood is a marginal, $E\ L_p^2$ does not depend on ϕ.

(b) $H_\theta^\phi \overset{L}{=} N(\phi, \theta_0-\theta) \overset{w}{\to} N(\phi_0, \theta_0-\theta) \overset{L}{=} H_\theta^{\phi_0}$ as $\phi\to\phi_0$.

(c) H_θ^ϕ have continuous distribution functions, so all intervals are continuity intervals.

(d) $f(x; \theta,\phi)$ is continuous in ϕ.

Further, the estimator s_{nk}^2 meets the Type I and Type II lower bound by Lemma 1.6.10, and so is fully effi-

cient (Φ_B and P_B).

6.9.2 Model B. The issues of Theorem 6.8.2 are avoided by noting that for any $\phi^{(\infty)}$ in Φ_B, the nuisance eliminating distributions can be structured to have support in a compact set Φ_0. (See Section 5.11.) Hence, for the restricted parameter space Φ_0 Theorem 6.5.5 implies that κ_p^* may be used for a lower bound for estimators which are consistent for all nuisance parameters in Φ_0^∞. This family of estimators contains all estimators consistent on Φ_B.

It is true that lower bound for both Type I and Type II models is met by the CMLE--however, the current proof requires substantial use of the theory of large deviations and so is omitted.

6.9.3 Model C. Here we have only a Type II model lower bound. It has not been established that the CMLE meets that bound.

6.9.4 Model D. The density $f(x;\theta,\phi)$ is not continuous in ϕ. Thus Theorem 6.8.2 may not be used. Thus the bound has not been established for estimators consistent on Φ_B. The MLE $\bar{Z}_{nk}$, however, is consistent for all sequences in R^∞ and meets the lower bound for estimators consistent

over all such sequences by Lemma 1.6.10. Further, it meets the Type I lower bound for aev for all null sequences with only a finite number of different values. (See discussion following Equation (8).) It meets the Type II lower bound for $\Theta \times P_B$.

6.9.5 Model E. As the space Φ is compact, there is no problem with null sequences. The estimator $\overline{W}$ is fully efficient aev as an estimator of the transformed parameter $\frac{1}{2}-2\theta^2$ by Lemma 1.6.10.

REFERENCES

Andersen, E. B. 1970. Asymptotic properties of conditional maximum likelihood estimators. *J. R. Statist. Soc. B* 32: 283-301.

Andersen, E. B. 1973. *Conditional inference and models for measuring.* Copenhagen: Mentalhygiejnisk Forlag.

Bahadur, R. R. 1960. Asymptotic efficiency of tests and estimates. *Sankhyā* 22: 229-252.

Bahadur, R. R. 1964. On Fisher's lower bound to asymptotic variances. *Ann. Math. Statist.* 26: 139-142.

Bahadur, R. R. 1967. Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* 38: 303-324.

Bahadur, R. R. 1971. *Some Limit Theorems in Statistics.* Philadelphia: SIAM.

Barankin, E. 1949. Locally best unbiased estimates. *Ann. Math. Statist.* 20: 447-501.

Barnard, G. A. 1963. Some logical aspects of the fiducial argument. *J. R. Statist. Soc. B* 25: 111-114.

Barndorff-Nielsen, O. 1973. On M-ancillarity. *Biometrika* 60: 447-455.

Barndorff-Nielsen, O. 1976. Plausibility inference (with discussion). *J. R. Statist. Soc. B* 38: 103-32.

Basu, D. 1977. On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* 72: 355-367.

Bickel, P. J. and Doksum, K. A. 1977. *Mathematical statistics.* San Francisco: Holden-Day.

Billingsley, P. 1968. *Convergence of probability measures.* New York: Wiley.

Chapman, D.G. and Robbins, H. E. 1951. Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.* 22: 581-586.

Chung, K. L. 1974. *A course in probability theory.* 2nd ed. New York: Academic Press.

Cox, D. R. 1958. Some problems connected with statistical inference. *Ann. Math. Statist.* 29: 357-372.

Cox, D. R. 1972. Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* 34: 187-220.

Cox, D. R. 1975. Partial likelihood. *Biometrika* 62: 269-276.

Cox, D. R. and Hinkley, D. V. 1974. *Theoretical statistics.* London: Chapman and Hall.

Dawid, A. P. 1975. On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. *J. R. Statist. Soc. B* 37: 248-258.

Efron, B. 1975. Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.* 3: 1189-242.

Feller, W. 1957. *An introduction to probability theory and its applications I.* 2nd ed. New York: Wiley.

Fisher, R. A. 1935. The logic of inductive inference. *J. R. Statist. Soc.* 98: 39-82.

Kalbfleisch, J. D. and Sprott, D. A. 1970. Application of likelihood methods to models involving large numbers of parameters. *J. R. Statist. Soc. B* 32: 175-208.

Kiefer, J. 1952. On minimum variance unbiased estimators. *Ann. Math. Statist.* 23: 627-29.

Kiefer, J. and Wolfowitz, J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27: 887-906.

Kullback, S. 1968. *Information Theory and Statistics.* Dover, New York.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. *Ann. Math. Statist.* 22: 79-86.

Johnson, N. L. and Kotz, S. 1970. *Continuous univariate distributions-I, distributions in statistics.* New York: Wiley.

LeCam, L. 1953. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Statist.* 1: 277-329.

Lehmann, E. L. 1959. *Testing statistical hypotheses.* New York: Wiley.

Loeve, M. 1977. *Probability theory I.* 4th ed. New York: Springer-Verlag.

Neyman, J. and Scott, E. L. 1948. Consistent estimates based on partially consistent observations. *Econometrika* 16: 1-32.

Plackett, R. L. 1977. The marginal totals of a 2x2 table. *Biometrika* 64: 37-42.

Rao, C. R. 1963. Criteria of estimation in large samples. *Sankhyā A* 25: 189-206.

Rao, C. R. 1973. *Linear statistical inference and its applications*. 2nd ed. New York: Wiley.

Rudin, W. 1964. *Principles of mathematical analysis*. 2nd ed. New York: McGraw-Hill.

Sprott, D. A. 1975. Marginal and conditional sufficiency. *Biometrika*, 62: 599-605.

Stein, C. 1964. Inadmissibility of the usual estimator of the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.* 16: 155-160.

Widder, D. V. 1941. *The LaPlace transform*. Princeton: Princeton University Press.

VITA

Bruce George Lindsay was born on March 7, 1947, in The Dalles, Oregon, to George S. and Geneva E. Lindsay. He graduated from The Dalles High School in 1965. Four years later he graduated from the Honors College of the University of Oregon with a Bachelor of Arts degree in Mathematics. That same year, 1969, he married Teresa Ann Goff, also of The Dalles. The school year 1969-70 started in graduate study in mathematics at Yale University, but ended with enlistment in the U.S. Coast Guard. The bright spot of four years service therein was the birth of a son Dylan in 1972. Since 1974 Bruce has been enrolled as a Ph.D. student in the Biomathematics program at the University of Washington.