FINITE-SAMPLE PROPERTIES OF MAXIMUM-LIKELIHOOD ESTIMATORS

by

Alex McMillan

A thesis submitted to the Faculty of Graduate Studies and Research, in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics,
McGill University,
Montreal.

September, 1978.

FINITE-SAMPLE PROPERTIES OF MAXIMUM-LIKELIHOOD ESTIMATORS

by

Alex McMillan

Abstract

Finite-sample properties of maximum-likelihood estimators are presented; the discussion is largely confined to properties which apply to parametric models with identically and independently distributed observations, and where the whole parameter is of interest. We first review various criteria which have been advanced to ensure that the estimator is specific for the parameter and is close to it. This is followed by an account of invariance and of conditions for sufficiency. The survey is rounded out by a discussion of conceptual and computational problems encountered in maximum-likelihood estimation and of the problem of making probability-type statements using the maximum-likelihood estimator.

Department of Mathematics,
McGill University,
Montreal.

M.Sc.

September, 1978.

FINITE-SAMPLE PROPERTIES OF MAXIMUM-LIKELIHOOD ESTIMATORS

par

Alex McMillan

Résumé

Ce mémoire présente des propriétés non-asymptotiques de l'estimateur du maximum de la vraisemblance. La discussion se limite en grande partie aux propriétés applicables à des modèles paramétriques avec observations distribuées indépendamment et identiquement, où l'on s'intéresse à toutes les composantes du paramètre. On passe d'abord en revue les critères qui assurent la justesse de l'estimateur et mesurent son éloignement du paramètre. Il suit un exposé de l'invariance et des conditions qui assurent la suffisance de l'estimateur. Nous terminons par une discussion des problèmes conceptuels et numériques qui entourent l'application de la méthode, ainsi que du problème de son utilisation dans des énoncés à charactère probabiliste.

Département de Mathématiques,
McGill University,
Montréal.

M.Sc.

Septembre, 1978.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor George P.H. Styan for suggesting the topic and for advising me through the course of my research. I also wish to thank him very warmly for encouraging me to return to graduate studies.

It is a pleasure to acknowledge discussions with Louis-Paul Rivest which led to the counterexample in Subsection 2.1.4. I am also thankful to Professors Timo Mäkeläinen, Klaus Schmidt and George P.H. Styan for making available a preliminary revision of their 1976 discussion paper.

Finally, I would like to thank Marlene Pyykko for her patience in skillfully and swiftly typing almost all the thesis. I am also grateful to Patricia Babecki for additional typing assistance.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

SECTION 1.1 POINT ESTIMATION

The problem of parametric point estimation is one where a set of data $x = (x_1, \ldots, x_m)$ (which can be regarded as a point in m-dimensional Euclidean space) is taken to be the realization of a random variable whose distribution, though unknown, can be assumed to belong to a certain parametric family, so that one can denote the probability that

$$x_1 \leq x_{o_1},\ x_2 \leq x_{o_2},\ \ldots,\ x_m \leq x_{o_m},$$

or $x \leq x_o$ for short, by:

$$P(x \leq x_o) = F(x_o;\ \theta)\ .$$

The functional form $F(\cdot;\ \cdot)$ of the multivariate cumulative distribution function is completely specified. However, it can only be said of θ that its value lies in some subset Θ of k-dimensional Euclidean space. It is the purpose of parametric point estimation to find, from the observed sample x, a function $\hat{\theta}(x)$ of the data (termed an estimator) which ranges over Θ and which can be said to represent a reasonable value for the unknown parameter θ.

Although some have criticized the use of the term 'estimator', preferring to speak only of 'estimates' (Fisher, 1958, p.7), we note that everyday language refers to an estimate as a definite valuation of

a quantity i.e. as a specified number, and it seems best to us to distinguish the estimator, as the agent or process of estimation, from its product, the estimate (Carnap, 1962, p.524).

Situations where Θ can be narrowed down to a finite collection of points will not interest us, since the approach to such multiple decision problems is essentially different from the one taken here. Rather, the cardinality of Θ will be that of the set of real numbers. In order to make point estimation meaningful, it must also be assumed that the parameter is identifiable, that is, that no two distinct values θ_1 and θ_2 of Θ are such that $F(\cdot;\ \theta_1) \equiv F(\cdot;\ \theta_2)$.

Justification for Point Estimation

Before going further, it may be noted that some have questioned the legitimacy of point estimation as a statistical problem. Point estimation is a rather weak form of inference: it merely produces a single number $\hat{\theta}$ and does not allow us to make probability-related statements such as are available in tests of significance. There is some feeling that the data cannot possibly be reduced to a single number, and that one should look to interval estimation as the appropriate alternative. The argument is advanced (from a Bayesian viewpoint) by Tiao & Box (1973) that if a client wants a single number (a point estimate) as the outcome of the statistical inference, he ought to be told that he shouldn't get it.

We will not attempt to answer such critiques directly; indeed, the variety of criteria which have been seriously proposed for point estimation testifies to the essential 'fuzziness' of the process. We will,

however, illustrate a few practical situations which either demand a point estimate or which are such that the analysis is considerably enhanced by the possibility of point estimation. (Questions relating to the making of probability-like statements will be discussed further in Section 4.3.)

In land surveying, several measurements of a quantity (say, the distance between two points A and B) may be available, and past experience may suggest a reasonable parametric family to represent this problem. Here, presumably, it will do no good to say that the surveyor should be satisfied with the fact that the distance is probably between 1.71 km and 1.73 km. The purpose of the survey is to construct a map, and sooner or later the measurements will have to be integrated with others and the distance between A and B will then be reduced to a point estimate: the nature of the medium allows no other reasonable conclusion.

Now, there is a very wide class of situations where the parameter θ is in a sense 'real', while the observations can be construed to have been developed specifically to lead to an assessment of the parameter. In the land-surveying example, the basic observations might be those of readings of angles, or of the timing of radar pulses, or of distance on a photograph, etc. Where the parameter is a physical quantity, point estimation has, we feel, a clear reason for existence. It is justified by our need to know some property of a real object. On the other hand, many situations are not of the 'measurement' type, but rather are a form of 'modeling' of the problem by means of a parametric family. To take an example from Hartigan (1967) (who discusses another aspect of the

'measurement-modeling' dichotomy) an experiment set up to evaluate the temperature of the sun is of what we call the measurement type. In contrast, when a series of temperatures are taken at various times, it may be desirable to interpret the series by means of a stochastic model which might involve, say, a parameter representing the mean temperature and its variance. Philosophically, the nature of the parameter here is different, it is a construct of our minds more than a real object. Yet even when the parameter is purely conceptual and not a physical quantity, it may be extremely useful to have a point estimate for the parameter. An example may be taken from the history of Genetics.

Around 1913 it had already been established that the genetically-determined traits of an individual are determined by pairs of genes (presumably located on chromosomes), one member of the pair representing the paternal inheritance and the other, the maternal. Simple genetic traits can take two characters or aspects (such as white or red eye pigmentation in fruit flies) denoted by A and a, B and b etc. A dominant character A is expressed whenever the individual's gene pair is AA, Aa or aA, while the recessive character a is taken to result from the presence of the doubly recessive gene pair aa. It was an object of investigation to study the frequency of offsprings with expressed character pairs AB, Ab, aB and ab resulting from the mating of pure hybrid parents whose gene pairs are of the form AaBb, the probabilities of each character pair being denoted p_{AB}, p_{Ab}, p_{aB}, p_{ab}. With genetic theory already asserting that $p_{AB} + p_{Ab} = p_{AB} + p_{aB} = 0.75$, the above probability distribution can be parametrized by the cross-over frequency $\theta_{AB} = p_{Ab} + p_{aB}$. Now why should one be interested in point estimates for θ_{AB} ?

The geneticist A.H. Sturtevant (1891-1970) interpreted θ_{AB} as a (conceptual) distance between the loci of the genes for trait pairs (A,a) and (B,b). The availability of estimates for several such trait pairs enabled him to determine that for many triplets A,B,C with θ's sufficiently small, one observed an almost linear relationship $\hat{\theta}_{AB} + \hat{\theta}_{BC} = \hat{\theta}_{AC}$. From this last observation it was deduced that genes are arranged linearly on the chromosome. It then became possible to prepare 'gene maps' showing relative conceptual distances between gene loci. Such maps have been a great conceptual aid in manipulating data in the exploration of genetic phenomena, and only point estimates can make them possible. (The above account is abstracted from Whitehouse, 1969, pp.77-90.)

Justification for point estimation could also be made on the basis of Decision Theory. However, we feel point estimates are not really decisions about the value of θ, although there is some controversy attached to this view. (Tukey, 1960, provides a lucid argument against it; the distinction between estimation and decision is supported, e.g., in Fisher, 1959, p.100; Hacking, 1965, p.164; Plante, 1971.) Although Decision Theory provides very useful criteria for point estimators, we take the view that point estimation is a legitimate, if primitive, form of scientific inference, and that a wider class of criteria is to be considered.

SECTION 1.2 MAXIMUM LIKELIHOOD ESTIMATION

We will now restrict our attention to parametric families $\{F(\cdot; \theta) : \theta \in \Theta\}$ where all members $F(\cdot; \theta)$ are absolutely continuous with respect to a single σ-finite measure λ. In other words, it will be assumed that the members of the parametric family can be represented by a set of densities

$$p(x; \theta) = \frac{dF(\cdot; \theta)}{d\lambda}(x) .$$

The dominating measure λ is usually the Lebesgue measure, in which case we speak of continuous observations, or a counting measure, when the observations are said to be discrete. While these two measures are the most common, it is sometimes necessary to use some other measure. For example, Proschan & Sullo (1976) use as a dominating measure the mixture of the Lebesgue measure on n-dimensional Euclidean space R^n, and of the Lebesgue measures on sets of the form $\{x : x_i = x_j\}$ and on all the intersections of such sets (e.g. $\{x : x_i = x_j = x_k\}$) considered as subsets of R^{n-1}, R^{n-2}, ..., R^1. It should be noted that while we may conceive of the data as consisting of continuous observations, the observations that are available are always essentially discrete, and that the use of a family of absolutely continuous distributions involves some approximation. The observations may be assumed continuous to permit the use of a simple parametric family, and because they represent quantities - such as distances - which in various experiments could be evaluated with varying degrees of precision. The method of maximum likelihood estimation uses the quantity $p(x; \theta)$ with x fixed at the observed value;

$p(x; \theta)$ is called the *likelihood function* and it is defined over the parameter space Θ. The maximum-likelihood estimator $\hat{\theta} = \hat{\theta}(x)$ of θ is usually defined as that value $\theta \in \Theta$ at which the supremum of the likelihood function is attained, i.e.,

$$p(x; \hat{\theta}) = \sup\{ p(x; \theta) : \theta \in \Theta \} .$$

Throughout this thesis the maximum-likelihood estimator will be abbreviated MLE. Another definition of MLE is sometimes used: with $\nabla = \nabla_\theta$ denoting the gradient operator with respect to θ, the MLE is defined as the solution of the likelihood equation

$$\nabla_\theta \log p(x; \hat{\theta}) = 0 .$$

When there are several solutions to the likelihood equation, that solution which maximizes $p(x; \theta)$ is taken as $\hat{\theta}$. Usually the two definitions coincide, but the second one is sometimes adopted, for example, in troublesome cases where the likelihood function has a singularity on the boundary of the parameter space.

History

The terms 'likelihood function', 'maximum likelihood estimation', etc., originate with Fisher (1922). There is some dispute as to whether Fisher originated the concept, which has been attributed variously to F.Y. Edgeworth, C.F. Gauss, P. Laplace, D. Bernoulli, and others. Some references to the history of MLEs are Edwards, 1974, Kendall, 1961, and Pratt, 1976. See also the survey by Norden (1972; 1973).

The issue is clouded by the fact that MLEs can be regarded as a particular case of 'Bayes' estimators. If the parameter is regarded as a random variable with distribution having density $\pi(\theta)$, it is possible

to interpret $p(x;\ \theta)$ as the density of the distribution of x conditional on θ, and this allows us to use Bayes' formula to derive the density of θ, conditional on $x = x_o$:

$$\pi^*(\ \theta \mid x = x_o\ ) = \frac{\pi(\theta)p(x_o;\ \theta)}{q(x_o)}\ ,$$

where $q(x_o)$ can be regarded as a normalizing constant. $\pi(\cdot)$ is termed the prior distribution and $\pi^*(\cdot)$, that of the posterior distribution of θ. A Bayes estimator is defined by taking some appropriate characteristic of the posterior, such as its mean, median or mode. Now if $\pi(\cdot)$ is the uniform distribution on Θ, the MLE corresponds to the Bayes estimator using the mode of the posterior. Whatever the resolution of the debate over priority for the invention of MLEs, it may be said that Fisher was instrumental in developing most of the basic theoretical properties of MLEs, in such fundamental papers as Fisher (1922, 1925, 1934). Further historical notes will accompany the discussion of specific properties.

Motivation for the use of MLEs

In the body of this thesis we will be concerned with properties of MLEs, and the overall performance of the estimator with respect to those properties can be taken as a justification for its use. It may, however, be of some interest to try to motivate the use of MLEs from general considerations.

The correspondence between MLEs and Bayes estimators can provide one motivation for the use of the method, albeit a rather weak one since, when Θ is unbounded, no proper uniform density $\pi(\cdot)$ can integrate to

one. Another motivation with a Bayesian 'flavour' has been given by Higgins (1977). It relies on the idea that the posterior distribution can be used as the prior distribution for analyzing any other data set obtained under similar circumstances. Now for a given data set x, and for some prior distribution π whose density does not vanish on Θ, a posterior density π^* can be obtained. If π^* is now used as the prior distribution in conjunction with the same data set x, a second posterior density π^{**} can be obtained. The process can be repeated indefinitely, at each step using the posterior of the last step as the prior. It is easy to see that this iterated posterior density will converge to a degenerate distribution concentrated at the MLE.

Beyond such Bayesian-style justifications, it is tempting to see in the likelihood function a 'dual' of the density, so that just as knowing the density explicitly enables us to say something about the probable location of the variate x, so knowing the likelihood allows us to say something about the 'likely' value of θ. The duality, however, is only approximate, since the behaviour of the two functions $p(\cdot; \theta)$ and $p(x; \cdot)$ is quite different in general.

Models

It is convenient to speak of the model as the conjunction of the data x (considered as a random variable), of the functional form of the densities $p(x; \theta)$, of the range Θ of distributions in the parametric family, and possibly of other factors which may be relevant to the situation at hand (such as the presence of a reasonable prior distribution).

An important type of model is one where the data x consists of n jointly (possibly multivariate) observations $x = (x_1, \ldots, x_n)$, each observation having the same distribution. If the density of that common parent distribution is denoted $q(\cdot; \theta)$ and if the density of x is denoted $p(\cdot; \theta)$, we have that the likelihood function can be expressed as

$$p(x; \theta) = \prod_{i=1}^{n} q(x_i; \theta) .$$

While we will primarily be interested in models where the observations are taken to be independently and identically distributed, we will nevertheless speak of p as being the density (instead of the more standard usage, where density always refers to the parent distribution). This convention allows a more general nomenclature to be followed.

SECTION 1.3 PURPOSE AND NOTATION

The properties of MLEs most often mentioned are asymptotic ones. This means that limiting properties of the sequence of MLEs $\{\hat{\theta}_n\}$ are studied, where $\hat{\theta}_n$ is the MLE (considered as a random variable) for a sample of size n in a model with identically and independently distributed observations. For example, it can be established (under regularity conditions) that the sequence $\{\hat{\theta}_n\}$ tends in distribution to the degenerate distribution concentrated at θ. The main argument against considering such asymptotic properties is that they can guarantee nothing about the behaviour of the estimator for any given sample size n. In other words, if we consider another sequence $\tilde{\theta}_n$, such that for $n < 10^{1000}$, $\tilde{\theta}_n = 0$ and for $n \geq 10^{1000}$, $\tilde{\theta}_n = \hat{\theta}_n$, then the sequence $\{\tilde{\theta}_n\}$ enjoys all the asymptotic properties enjoyed by $\{\hat{\theta}_n\}$, while $\tilde{\theta}_n$ is completely useless in practice. (Savage, 1976, p.453; see also: Kallianpur & Rao, 1955; Fisher, 1959, p.146). On the other hand (as noted by Chernoff, 1976), the theory of MLEs for finite samples is often difficult or impossible to formulate. The asymptotic theory does say something about one aspect of the sequence $\{\hat{\theta}_n\}$ as a whole, and it can be hoped that when there are no sudden 'breaks' in the sequence (as in $\{\tilde{\theta}_n\}$ above), the asymptotic property is reflected in weaker form in individual members $\hat{\theta}_n$ of the sequence.

In this thesis we will survey non-asymptotic properties of maximum likelihood estimation. While we will avoid properties of the sequence of estimators, it will not be possible to restrict ourselves to properties of estimators $\hat{\theta}$ per se. Some of the more significant properties of maximum likelihood estimation pertain, not to $\hat{\theta}$ itself, but to the

manner in which it is produced. Thus when we say that MLEs are invariant to reparametrizations (so that if $p(x; \theta) \equiv q(x; \phi)$, with $\phi = \psi(\theta)$, then $\hat{\phi} = \psi(\hat{\theta})$), we are really saying something, not about estimators $\hat{\theta}$ or $\hat{\phi}$, but about the principle of estimation which allows these two estimators to be related in this manner.

Although it often happens that only a subparameter θ_1 of θ is of interest (with the remaining part of θ being termed an incidental or nuisance subparameter), we will only consider criteria and properties which apply to situations where the whole parameter is of interest. If one seeks to motivate the use of a method on the basis of 'philosophical' considerations such as the apparent duality of the density and the likelihood, it would appear desirable to use some modification of the likelihood function instead of the likelihood function itself. Kalbfleisch & Sprott (1970) explore several possible modifications.

An excellent survey of maximum-likelihood estimation was made by Norden (1972; 1973), with an emphasis on asymptotic properties, however.

Plan

In Chapter 2 of this thesis, we will consider the performance of the MLE relative to criteria which ensure that an estimator does, in a sense, estimate the parameter; and relative to criteria which measure how close an estimator comes to the true value. In Chapter 3, we review the sufficiency of the MLE and various invariance properties of the method of maximum likelihood. In Chapter 4, we take a brief look at some of the factors which hinder the applicability of the method. Chapter 5 contains a summary of properties of maximum-likelihood estimation. The three appendicies will be refered to at appropriate places in the main portion of the thesis.

Notation

- The probability of an event A will be denoted $P(A|\theta)$ or $P(A)$, according to whether or not it is necessary to specify the parameter of the probability distribution.
- Both the density and the likelihood function will be denoted $p(x; \theta)$ or $q(x; \theta)$, it being clear from the context whether x or θ is regarded as fixed.
- Likewise, no distinction will be made between x regarded as a random variable and x as its realization. Again, the context will specify which usage is the appropriate one.
- The parameter space will generally be denoted Θ while the space over which the random variable x takes values will be denoted Ξ.
- Expectation and variance will be denoted E_θ or E, V_θ or V.
- Parameters will be denoted by lower case Greek letters θ, ϕ, etc., and their estimators will be denoted $\hat{\theta}$, $\tilde{\theta}$, $\hat{\phi}$, $\tilde{\phi}$, etc.
- Parameters and their associated functions are generally to be regarded as vectors. This is so commonly the case here that no special notation will usually be adopted to distinguish scalar from vector parameters. The same remark holds true of the data x, which is usually a vector. An exception will be in denoting the parameter and variate of a multivariate Gaussian (or normal) distribution, where vectors will be denoted by wavy underlining.
- The matrix (or vector) of derivatives of a function f is denoted

$$\nabla_\theta f = \nabla f = \left(\frac{\partial f_j(\theta)}{\partial \theta_j} \right),$$

and the matrix of second derivatives of a scalar-valued function f will be represented by

$$\nabla\nabla^T f = \left(\frac{\partial^2 f}{\partial \theta_i \partial \theta_j} \right).$$

(Superscript T denoting transpose.)

It will often be necessary to specify the carrier or support set of a function. This will be done by means of a pair of pointed brackets to denote the indicator function of the logical statement they enclose. Thus $\langle\, a \in A \,\rangle$ denotes the function whose value is one when a is a member of A, and zero otherwise. Likewise

$$\langle\, a \leqslant \theta \leqslant b \,\rangle = \begin{cases} 1 & \text{if } \theta \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

(Our usage of $\langle\, , \,\rangle$ should not be confused with a more usual usage, where $\langle X \rangle$ denotes EX, the expectation of the variate X, e.g., Barnard, 1973.)

CHAPTER 2 SPECIFICITY AND CLOSENESS CRITERIA

SECTION 2.1 SPECIFICITY

2.1.1 INTRODUCTION

The definition of estimator alluded to in the first chapter is rather incomplete since it is based on the "appearance" of the estimator: it has merely been required that the statistic $\hat{\theta}$ range over a subset of Θ - the question of $\hat{\theta}$ falling not in Θ but on the boundary of Θ will be considered in Subsection 4.1.1.

The principal requirement we would like an estimator $\hat{\theta}$ to satisfy is that of being close, in some sense, to the true 'target' value of θ. Several closeness criteria will be considered in the next section; for most criteria, however, we are faced with a difficulty which arises from the inability, in the Frequency outlook, of averaging the performance of an estimator over the class of possible populations or equivalently, over the parameter space. Such a difficulty disappears when a prior distribution is available. Consider as a trivial example the problem of estimating the true proportion π from a sample of n Bernoulli trials. The estimator $\tilde{\pi} \equiv 0.5$ is in many senses a poor estimator: it totally rejects whatever information the sample might provide. However, if the true value of π were equal to 0.5, $\tilde{\pi}$ would undeniably be the closest estimator according to any reasonable criterion, for the precise population having a parameter value equal to 0.5. Even for populations whose parameter is only nearly equal to 0.5, $\tilde{\pi}$ may still be better than some quite reason-

able estimators. For instance, when closeness is measured by the average of the squared deviation of the estimate from the true value, i.e., $E(\tilde{\pi} - \pi)^2$, then $\tilde{\pi} \equiv 0.5$ is better than the usual estimator, the MLE $\hat{\pi}$ = "number of successes"/n over the whole range of π from $\frac{1}{2} - \frac{1}{2\sqrt{(n+1)}}$ to $\frac{1}{2} + \frac{1}{2\sqrt{(n+1)}}$; when $n = 3$ the trivial estimator is better than the MLE on the range $(0.25, 0.75)$ (Silverstone, 1957).

In order to exclude such extremely partial estimators as $\tilde{\pi}$, one could try to formulate a criterion of impartiality. Such a criterion would then have to be sufficiently weak so as not to demand that the estimator have exactly the same performance for all populations in Θ: it would seem unreasonable to demand that the distribution of any estimator be of the same 'shape' for all values of θ.

A slightly different approach is to ask that the estimator, or rather the estimation method, be 'specific' to the parametrization in the broad sense that when the same data are used to estimate θ and, separately, to estimate $\psi(\theta)$ (a non-trivial transformation of θ), then one would not want the estimator of $\psi(\theta)$ to equal the estimator of θ. We would like to assure ourselves that $\hat{\theta}$ is 'targeted' to θ but not to $\psi(\theta)$, that it is on the correct 'scale' ('scale' is used here in the wide sense of 'logarithmic scale', 'harmonic scale', etc., not merely in the sense of a unit of measurement). For the nonce we will use the term 'specificity' to denote the fact that an estimation method produces estimators which are 'on target', or specific. The word 'consistency' would perhaps be better suited but its usual meaning (i.e. probability consistency) is so well entrenched that confusion is best avoided.

2.1.2 COGREDIENCE-SPECIFICITY

A first approach to the question of specificity uses the notion of cogredience which we will discuss at length in Chapter 3. Briefly, a model has a cogredience structure when one can find a group G of transformations on the space of observations and an (induced) group $\bar{G}$ of transformations on the parameter space, so that to every transformation $g \in G$, one can find $\bar{g} \in \bar{G}$ to satisfy $p(g(x); \bar{g}(\theta)) \equiv p(x;\theta)$ for all θ and almost all x. An estimator t is 'cogredient' when $t(x) = \hat{\theta}$ and $t(g(x)) = \bar{g}(\hat{\theta})$ for almost all x. The suggestion to use a specific type of cogredience to arrive at a notion of specificity has been made by Barnard (1962) (using scale invariance) and, in a more general setting by Lehmann (1950a, pp.1-17; 1959, p.10). In this section we will primarily give an illustration of the principle by justifying the specificity of a competitor to the MLE. Consider the problem of making inferences about the mean $\underset{\sim}{\mu}$ of a k-variate Gaussian distribution with known scalar covariance matrix of the form $\sigma_o^2 I$, on the basis of an observation $\underset{\sim}{x}$, when $k \geq 3$. The problem has at least the following symmetry: if $\underset{\sim}{x}$ were to be measured on a different linear scale (say that instead of measuring a length in centimetres, one measured it in inches) then instead of reporting an observation $\underset{\sim}{x}$ one would report $\underset{\sim}{y} = a\underset{\sim}{x}$, where a is the scaling factor ($a = 0.3937$ approximately in converting from centimetres to inches). The variate $\underset{\sim}{y}$ is not distributed as Gauss $(\underset{\sim}{\mu}, \sigma_o^2 I)$ but another member of that same parametric family can be found, namely Gauss $(a\underset{\sim}{\mu}, a^2\sigma_o^2 I)$, which describes the distribution of the new variate $\underset{\sim}{y}$. We might therefore agree that an estimator $\underset{\sim}{t}(\underset{\sim}{x}) = \underset{\sim}{\tilde{\mu}}$ is specific in the sense of being scale invariant, if for all a, the estimator yields a

value $\underset{\sim}{t}(a\underset{\sim}{x}) = a\tilde{\underset{\sim}{\mu}}$ (as compared to the true parameter $a\underset{\sim}{\mu}$ when measured in the appropriate units) when the data are reported as $a\underset{\sim}{x}$ and the variance is $a^2\sigma_o^2 I$. In symbols,

$$(\underset{\sim}{x}, \underset{\sim}{\mu}, \sigma_o^2, \tilde{\underset{\sim}{\mu}}) \rightarrow (a\underset{\sim}{x}, a\tilde{\underset{\sim}{\mu}}, a^2\sigma_o^2, a\underset{\sim}{\mu}) .$$

Thus in this problem (as enlarged by the requirement of scale invariance or covariance) an estimator of the type considered by James & Stein (1961):

$$\tilde{\underset{\sim}{\mu}} = \left[1 - \frac{(k-2)}{n} \frac{(a\sigma_o)^2}{\|\underset{\sim}{y}\|^2} \right] \underset{\sim}{y} \qquad (1)$$

is at least scale-invariant, so that it can be claimed to be specific for $\underset{\sim}{\mu}$.

We note that there seems to be some degree of arbitrariness in the selection of the group of transformations under which the estimator is to be cogredient. In the above example, it might be natural to impose cogredience with respect to all affine transformations, i.e.,

$$(\underset{\sim}{x}, \underset{\sim}{\mu}, \sigma_o^2, \tilde{\underset{\sim}{\mu}}) \rightarrow (a\underset{\sim}{x} + \underset{\sim}{b}, a\underset{\sim}{\mu} + \underset{\sim}{b}, a^2\sigma_o^2, a\tilde{\underset{\sim}{\mu}} + \underset{\sim}{b}).$$

It is obvious that the estimator (1) is no longer cogredient under this wider group. Thus when several types of symmetry are possible for a model, conflicting criteria of 'cogredience-specificity' will result. For this reason it is perhaps best not to generalize the criterion beyond scale invariance; scale invariance also bears a strong relationship to the notion of physical dimensionality of a parameter, which notion should usually be respected.

Finally, we note from the result in Chapter 3, that the MLE is

always cogredient whenever there exists a cogredience structure in the model. *Therefore, the MLE will always be cogredience-specific.*

2.1.3 CRITERIA RELATED TO THE USE OF LOSS FUNCTIONS

Even before the formal introduction of decision theory, loss functions were being used in estimation theory to provide closeness criteria. In this subsection we will consider a class of specificity criteria which, though not necessitated by a specific loss function, are nevertheless closely associated with, and can be derived from, a loss function. Lehmann (1959, p.11), in a decision-theoretic setting, defines an estimator $\hat{\theta}$ to be (loss-) unbiased if:

$$E_\theta loss\ (\hat{\theta};\ \theta') \geq E_\theta loss\ (\hat{\theta};\ \theta)$$

for all $\theta' \in \Theta$. That is to say, an unbiased estimator attains minimum risk at the true value θ. Two instances of loss-unbiasedness which have rather wide currency in the statistical literature are: mean-unbiasedness (usually refered to simply as unbiasedness), which is unbiasedness using uniform quadratic loss:

$$loss\ (\hat{\theta};\ \theta) = \|\theta - \hat{\theta}\|^2 ; \tag{2}$$

and median-unbiasedness, which is produced, when using a one-dimensional parameter, by the absolute-value loss:

$$loss\ (\hat{\theta};\ \theta) = |\theta - \hat{\theta}| .$$

There does not appear to be much discussion of the kind of unbiasedness produced when a non-uniform quadratic loss:

$$loss\ (\hat{\theta};\ \theta) = g(\theta)\ \|\theta - \hat{\theta}\|$$

is used instead of (2), the uniform quadratic loss, even though it would sometimes be more natural (particularly in reliability studies - Canfield, 1970) to use, when θ is one-dimensional, the loss:

$$loss\ (\hat{\theta};\ \theta) = \left| \frac{\hat{\theta}}{\theta} - 1 \right|^2 .$$

Mean-unbiasedness

Just as quadratic loss is by far the most widely discussed loss function, mean-unbiasedness is the premier example of a specificity criterion in Statistics. The criterion boils down to requiring that the expected value of the estimator exist and be equal to the true parameter. The criterion is associated with Gauss, although it has been remarked (Barnard, 1962; Sprott, 1978) that unbiasedness per se was introduced into the study of linear models by Markov. Gauss apparently specified error-consistency for estimators: that the estimator should yield the true parameter value when the errors are all equal to zero.

Whatever the historical background, the fact remains that mean-unbiasedness is both mathematically very tractable and statistically very restrictive. Apart from tractability, the best that can be said about mean-unbiasedness as a criterion is that it effectively does rule out extremely partial estimators. Other arguments that have been advanced in favour of the criterion are, in our opinion, far less compelling. In particular, there is a certain circular quality to the argument (cf. Haldane & Smith, 1956) that when a biased estimator is being used to

produce a large number of estimates which must afterwards be summarized by a mean or a median, that the mean or median will itself be seriously biased.

Among the most serious objections to the criterion we may note:

(1) That an estimation method producing mean-unbiased estimators in one parametrization will only produce unbiased estimators for affine transformations of the parameter, so that the criterion is only applicable where only affine reparametrizations, if any, are to be considered: it therefore clashes with the property of invariance to reparametrization which is considered in Subsection 3.2.2.

(2) That it excludes estimators which have no expectation, even when the source of the divergence is a class of events of negligible probability.

(3) That the criterion is sometimes totally inapplicable because no estimator exists which is unbiased in the mean (Ghosh & Singh, 1970; Wasan, 1970, p.109; Tusnády, 1968), while there are other cases where the unique unbiased estimator takes values outside the parameter space (Cox & Hinkley, 1974, p.253; Sprott, 1978; Wasan, 1970, p. 108).

Of the above objections, we view the lack of invariance as being the most crucial, because it places restrictions on the ultimate use to which the estimator will be employed; even when those uses can be foreseen, it is far too easy to fail to see that some non-linear transformation will destroy any advantage which might accrue from using an unbiased estimator.

One unfortunate example of a slip occasioned by this situation

ought to be sufficient to drive home the point. Wasan (1970, p.169) considers the estimation of the reliability in a Weibull model with known shape and threshhold - that is, the observations are essentially from an exponential variate (with unknown scale parameter) raised to a known power. The use of the MLE is criticized because the MLE is biased and because the reliability estimator $\hat{\rho}$ for a single component of a system will most often be used, raised to the power m, to estimate (or predict) the reliability ρ^m of a system with m identical components arranged in series. In the situation being considered, for a true value $\rho = 0.951$ of the reliability, the MLE $\hat{\rho}$ will have mean $E\hat{\rho} = 0.938$, and for $m = 10$ the true reliability will be $\rho^{10} = 0.605$ while $(E\hat{\rho})^{10} = 0.527$, a 13% relative error (here, though, $E(\hat{\rho}^{10})$ would be of interest).

It is proposed to remedy the situation by using the minimum-variance unbiased estimator $\tilde{\rho}$ in place of $\hat{\rho}$. Now simple consideration of Jensen's inequality will show that $\tilde{\rho}^{10}$ must be biased, and biased positively, so that where the MLE is unduly pessimistic about the reliability, the proposed remedy will on the contrary be unduly optimistic. The following table, based on numerical integration, compares the performance in terms of bias and root-mean-square error, for the MLE and its alternative, for selected values of m.

m	True value	MLE	MVUE	Bias MLE	Bias MVUE	RMSE MLE	RMSE MVUE
1	0.951	0.940	0.951	-0.011	0.000	0.034	0.027
10	0.605	0.560	0.624	-0.045	0.019	0.154	0.140
20	0.366	0.336	0.408	-0.030	0.042	0.154	0.162

(Details in Appendix A) The bias in $\tilde{\rho}^{10}$ is already roughly 2 percentage points, against over 4 percentage points on the conservative side

by the MLE. For $m = 20$, the bias in $\tilde{\rho}^{20}$ is higher than that in $\hat{\rho}^{20}$.

A final comment about this is that reliability is hardly the type of situation where one should specify a symmetric loss function. Canfield (1970) has made a similar point and has suggested using a piecewise quadratic loss.

As illustrated by this example, and as should be expected from an estimator which is totally invariant to reparametrization, *the MLE is in general not mean-unbiased for the particular parametrization under consideration.* The question of whether any parametrization exists, for which the corresponding MLE is unbiased, does not appear to have been approached except in the special case of models of the exponential family. There, Barton (1956) has noted that with the notation:

$$p(x;\ \theta) = \exp\ (\ t(x)^T \phi(\theta) + \beta(\theta) + g(x)\ )\ ,$$

the 'mean value' reparametrization $\theta \rightarrow \tau(\theta)$ such that:

$$\tau(\theta) = -\ (\ \nabla_\theta \phi(\theta)\ )^{-1} \nabla_\theta \beta(\theta)$$

or equivalently,

$$\tau(\theta) = E_\theta t(x)\ ,$$

will have an MLE which is unbiased for the corresponding parameter and will be such that its variance will attain the Cramér-Rao bound; from the argument it also follows that this is the unique parametrization (aside from affine transformations) for which the MLE can be unbiased.

The mean-value parametrization is not necessarily very convenient. For example, in the two-parameter univariate Gaussian with mean μ and

variance σ^2, the mean value parametrization is:

$$\tau(\mu, \sigma^2) = (\mu, \mu^2 + \sigma^2) .$$

Median-unbiasedness

Median-unbiasedness is the only other type of unbiasedness which has been given any consideration at length. From

$$\mu = \textit{median}\ x$$

$$\text{and} \quad \mu < a < b \quad \text{or} \quad \mu > a > b$$

$$\text{imply:} \quad E\,|x - a| \leq E\,|x - b| ,$$

(cf. Wasan, 1970, p.119) one can immediately deduce that when an estimator takes the median of its distribution at the true parameter value, then that estimator will have minimum risk at the true parameter value for the absolute value loss. Thus an estimator $\hat{\theta}$ is median-unbiased for θ when the median of $\hat{\theta}$ is θ.

Among other advantages of this criterion Birnbaum (1964) has noted the fact, easily verified, that if $\hat{\theta}$ is median-unbiased for θ and if g is a monotone function on Θ, then $g(\hat{\theta})$ will be a median unbiased estimator of $g(\theta)$. Thus, it can be claimed that this is a criterion which is capable of being applied where one may, at a later point, want to rescale the parameter, e.g., by taking its inverse, or its logarithm, etc.

Disadvantages of the criterion are that it does not appear to have a multi-dimensional analogue, and that when the observations have a discrete distribution, estimators based solely on the data will also have a discrete distribution where, typically, the median will only be known to lie

within a certain interval or a certain class. Birnbaum's suggestion (1964) that an estimator be made to have a continuous distribution by using randomization would seem rather artificial.

Because of the general invariance of the MLE to reparametrizations, we should again expect that *the MLE will not in general satisfy the median-unbiasedness criterion.* Thus when estimating the mean of a univariate Gaussian with known variance, the MLE will be median-unbiased; in other cases it is not: when estimating the variance of a Gaussian variate whose mean is known, the median of the MLE's distribution will always be above the true variance.

2.1.4 FISHER CONSISTENCY

Consistency as usually defined is an asymptotic property of a sequence of estimators: if the sequence converges in probability to the true parameter value as the sample size goes to infinity, the sequence (and hence, by a usual ellipsis, the estimator as a typical member of the sequence) is said to be consistent. Another definition of consistency which is not asymptotic, was often alluded to by Fisher (e.g., in 1922; 1959, p.144; 1935). Fisher gave a satisfactory definition for the discrete case and an adequate general definition was given by Kallianpur & Rao (1955). The asymptotic criterion is denoted probability consistency and the latter, Fisher consistency. Fisher consistency appears to be definable only when the observations are identically and independently distributed random variables, so that this situation will be assumed to the end of this subsection.

Discrete Case

We begin by considering the case of observations from a discrete variate taking value x with probability $p(x; \theta)$. Without loss of generality the x's can be relabeled to be 1, 2, 3... ; denote by $\underset{\sim}{p}$ the vector whose x-th component is $p(x; \theta)$. Let the number of observations of x (after relabeling) be n_x and denote by $\underset{\sim}{q}$ the vector whose x-th component is n_x/n, the observed proportion in the x-th cell. Both $\underset{\sim}{p}$ and $\underset{\sim}{q}$ take their values on

$$\Omega = [\,0,\,1\,] \times [\,0,\,1\,] \times \cdots$$

where $[0, 1]$ denotes the closed interval from zero to one and where the Cartesian product extends to k dimensions, or has countably-infinite dimensionality, according to whether only k values of x, or an infinite number of x's, are possible. An estimator $\hat{\theta}$ which can be written as

$$\hat{\theta} = H(\underset{\sim}{q})$$

$$\text{where } H : \Omega \to \Theta,$$

is Fisher-consistent if $H(\underset{\sim}{p}) = \theta$.

Fixed Carrier Case

In a more general setting, where the observations need not be discrete, the role of $\underset{\sim}{q}$ is taken by the empirical cumulative distribution function (ECDF):

$$F_x(\cdot) = F(y; x) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{k} \langle\, x_{i,j} \leq y_j \,\rangle$$

(where $x_{i,j}$ is the j-th component of the i-th observation) as a representation of the sample. Suppose therefore that the estimator $\hat{\theta}$ of θ is a functional of the ECDF, that is

$$\hat{\theta} = H(F_x) ,$$

where H is a mapping from an abstract space of distribution functions into Θ. The mapping H is then called a functional of its argument.

A simple example of a functional is:

$$H(F) = \int x \, dF(x)$$

(where x is one-dimensional) which yields the sample mean when F is the ECDF:

$$H(F) = \int x \, dF_x(x)$$

$$= \Sigma \, x_i \frac{1}{n} .$$

An estimator $\hat{\theta}$ is said to be Fisher-consistent for θ if:

$$H(F_\theta) = \theta \quad \text{for all} \quad \theta \in \Theta \tag{3}$$

for F_θ the multivariate cumulative distribution function:

$$F_\theta(x) = \int_{Q(x)} p(y; \theta) \, d\lambda(y) ,$$

$$Q(x) = \{ y \in \Xi : y_j \leq x_j \text{ for all } j \} .$$

Equation (3) means that the functional always yields the value θ when it is evaluated at the distribution indexed by θ.

As pointed out by Rao (1962a), the functional corresponding to the MLE is:

$$H(F) = \textit{mode}\ \{ \int \log p(x;\ \theta)\ dF(x)\ :\ \theta \in \Theta \}\ . \qquad (4)$$

Rao claims that with this functional the MLE is Fisher-consistent "without any restriction whatsoever". However, it would appear that some care is needed in using (4). First, the absolute maximum of the function inside the brackets is not necessarily taken at an interior point of Θ, nor is it necessarily taken at a unique value of θ, so that $H(F)$ is properly a set-valued functional ranging over subsets of Θ, including the null set (see Plante, 1976, for similar considerations). Second, we will demonstrate later on that the integral in (4) fails to be defined when F is the distribution function of some continuous models.

The first point is inessential inasmuch as we can agree, by convention (in the event that the distribution function F_θ should yield an H with several modes), that the criterion be satisfied so long as θ is one of those possible modes. Except for situations which give rise to an undefined integral, it would seem that the mode will be unique for a distribution function corresponding to a member of the family in the model under consideration.

Consider first the case where the family of densities has a common support or carrier set:

$$p(x : \theta) > 0 \quad \text{if and only if} \quad x \in S, \quad \text{for all} \quad \theta.$$

Define:

$$g(\theta;\ \theta_o) = \int_S p(x;\ \theta_o)\ \log p(x;\ \theta)\ d\lambda(x).$$

We wish to determine the mode or modes of $g(\theta;\ \theta_o)$, as θ ranges over Θ.

$$g(\theta;\ \theta_o) = -\int_S p(x;\ \theta_o)\ \log \frac{1}{p(x;\ \theta)}\ d\lambda(x)$$

$$= -\int_S p(x;\ \theta_o)\ \log \frac{p(x;\ \theta_o)}{p(x;\ \theta)}\ d\lambda(x)$$

$$+ \int_S p(x;\ \theta_o)\ \log p(x;\ \theta_o)\ d\lambda(x)$$

$$= g(\theta_o;\ \theta_o) - I(\theta_o;\ \theta)$$

where

$$I(\theta_o:\ \theta) = \int p(x;\ \theta_o)\ \log \frac{p(x;\ \theta_o)}{p(x;\ \theta)}\ d\lambda(x)$$

is the Kullback-Leibler Separator.

Now it is known that $I(\theta_o:\ \theta)$ is defined and positive for constant-carrier densities, except that $I(\theta_o:\ \theta)$ may be zero if $p(x;\ \theta) = p(x;\ \theta_o)$ almost everywhere. As this last contingency is excluded when the model is assumed to be indentified, we see that:

$$g(\theta_o;\ \theta_o) > g(\theta;\ \theta_o)$$

when $\theta_o \neq \theta$, so that the MLE is Fisher-consistent.

Variable Carrier Case

Now consider a family of densities with variable carrier:

$$S(\theta) = \{\ x\ :\ p(x;\ \theta) > 0\ \}\ .$$

For given θ_o and θ, let:

$$A = S(\theta) \cap S(\theta_o), \quad B = S(\theta_o) - S(\theta)\ .$$

Then:

$$g(\theta;\ \theta_o) = \int_{S(\theta_o)} p(x;\ \theta_o)\ \log p(x;\ \theta)\ d\lambda(x)$$

$$= \int_A p(x;\ \theta_o)\ \log p(x;\ \theta)\ d\lambda(x)$$

$$+ \int_B p(x;\ \theta_o)\ \log p(x;\ \theta)\ d\lambda(x). \qquad (5)$$

Strictly speaking the integral over B is ill-defined since $\log 0$ is not defined. However, one may use the convention $\log 0 = -\infty$, since this is the limit of $\log p$ as p tends to 0 from the positive side. It is, therefore, quite reasonable to set the second integral in (5) equal to $-\infty$ whenever $P(B|\theta_o) > 0$, and equal to zero when $P(B|\theta_o) = 0$. The function g is still well defined, provided the integral over A in (5) is finite or diverges to $-\infty$. When g is well defined, let:

$$q(x;\ \theta) = p(x;\ \theta)\ [\ P(A|\theta)\ ]^{-1}\ \langle\ x \in A\ \rangle\ ; \qquad (6)$$

then the integral over A in (5) is:

$$\int_A P(A|\theta_o)\ q(x;\ \theta)\ \log [\ P(A|\theta_o)\ q(x;\ \theta)\ ]\ d\lambda(x)$$

$$= P(A|\theta_o) \int_A q(x;\ \theta_o)\ \log q(x;\ \theta)\ d\lambda + P(A|\theta_o)\ \log P(A|\theta).$$

The family $\{\ q(\cdot;\ \theta)\ \}$ has constant support and $P(A|\theta_o) = P(\ S(\theta_o)|\theta_o\ ) = 1$, so that if:

$$\int_A q(x;\ \theta_o)\ \log q(x;\ \theta_o)\ d\lambda(x) < \infty\ , \qquad (7)$$

the integral over A in (5) is defined and

$$\int_A p(x;\ \theta_o)\ \log p(x;\ \theta)\ d\lambda(x) \leq \int_A p(x;\ \theta_o)\ \log p(x;\ \theta_o)\ d\lambda(x)$$

with equality if and only if:

$$\left.\begin{aligned} & P(\, S(\theta_o) \cap S(\theta) \,|\, \theta) = 1 \\ \text{and} \quad & \\ & \frac{p(x;\, \theta_o)}{P(S(\theta_o) \cap S(\theta) \,|\, \theta_o)} = \frac{p(x;\, \theta)}{P(S(\theta_o) \cap S(\theta) \,|\, \theta)} \text{ a.e. } \lambda \end{aligned}\right\} \tag{8}$$

The two conditions (8) together imply $p(x;\, \theta) = p(x;\, \theta_o)$ a.e. λ, which is excluded by the identifiability of the parameter. It therefore appears possible to write conditions for $g(\theta;\, \theta_o)$ to be defined in (5).

Most of the remainder of this subsection will be taken up with an example where (5) reduces to an undetermined form $g(\theta;\, \theta_o) = \infty / - \infty$. Before exhibiting the example, it should be mentioned that Fisher consistency could still be defined, even in such cases, by the further convention that, in taking the mode of $g(\cdot;\, \theta_o)$, those values of θ for which $g(\theta;\, \theta_o)$ is not defined are to be ignored. $g(\theta_o;\, \theta_o)$ will be defined (though it may be infinite) and it would appear, even when (7) fails to hold, that the number of θ for which $g(\theta;\, \theta_o)$ diverges will be at most countable. If this conjecture is verified, Fisher consistency could be said to hold in general for the MLE, albeit with appropriate conventions.

Ill-defined g

The following is an example where (5) cannot be defined. It was communicated privately to the author by Louis-Paul Rivest.

Consider first the density

$$q(x) = \frac{c}{x(\log x)^2} \quad \left(0 < x < \tfrac{1}{2} \right) .$$

where $c = \log 2$.

$$\int q(x) \log q(x)\, d\lambda(x) = -\int_0^{\frac{1}{2}} \frac{c}{x(\log x)^2} \log\left[x(\log x)^2\right] dx + \log c$$

$$= \int_c^{\infty} cy^{-1} - 2cy^{-2} \log y \, dy + \log c \qquad (9)$$

where the change of variables $y = -\log x$ was made. (9) diverges to $+\infty$ since:

$$\int_c^{\infty} y^{-2} \log y \, dy = \frac{\log c + 1}{c}$$

while $\int_c^{\infty} y^{-1} = +\infty$.

Now form the parametric location family with $\Theta = (-\infty, \infty)$ by setting:

$$p(x; \theta) = \tfrac{1}{2}q(x - \theta) + \tfrac{1}{2}q(x + \tfrac{1}{2} - \theta) .$$

Let $\theta = \theta_o + \frac{1}{2}$. Then we have:

$p(x; \theta_o)$ has support on $(\theta_o, \theta_o + \frac{1}{2}] \cup (\theta_o + \frac{1}{2}, \theta_o + 1]$

$p(x; \theta)$ has support on $(\theta_o + 1, \theta_o + \frac{3}{2}] \cup (\theta_o + \frac{1}{2}, \theta_o + 1]$

so that $B = (\theta_o, \theta_o + \frac{1}{2}]$, $A = (\theta_o + \frac{1}{2}, \theta_o + 1]$ and

$$\int_A p(x; \theta_o) \log p(x; \theta)\, dx = \int_{\theta_o + \frac{1}{2}}^{\theta_o} \frac{q(x - \theta_o - \frac{1}{2})}{2} \log \frac{q(x - \theta_o - \frac{1}{2})}{2} dx$$

$$= \int_0^{\frac{1}{2}} \frac{q(x)}{2} \log \frac{q(x)}{2} dx$$

$$= +\infty$$

while $P(B|\theta_o) = 1$ so:

$$\int_B p(x;\ \theta_o) \log p(x;\ \theta)\ dx = -\infty ,$$

and g in (5) is not well-defined.

Fisher Consistency in General

Although some problems crop up with the definition of Fisher consistency, the situation is rather satisfactory for the MLE. *Where the MLE can be defined by the 'mode' functional, the MLE is Fisher-consistent.* In particular the criterion is satisfied for discrete models and for continuous models with constant carrier. It is also satisfied, given some conventions, when the densities have variable carrier but are bounded in such a way that for every pair θ_o, θ, there exists a number $K(\theta_o, \theta) < \infty$ such that:

$$p(x;\ \theta) \leq K(\theta_o,\ \theta) ,$$

since this will allow (6) to be properly defined.

We note that other estimation methods also yield Fisher-consistent estimators. When the method of moments uses the sample moments $\Sigma x_i^m/n$, the moment-estimator will be Fisher-consistent. Ponnapalli (1976) shows Fisher consistency for a wide class of estimators in discrete models, under some regularity conditions on the densities. From his result, or directly, one can show that the minimum chi-squared and minimum modified chi-squared estimators are also Fisher-consistent. Other methods, such as the method of moments based on sample cumulants (k-statistics), are not Fisher-consistent.

2.1.5 SPECIFICITY CRITERIA: NOTES AND SUMMARY

The criteria considered above are far from exhausting the possibilities, although we believe the most common criteria have been covered. Here we will briefly consider a few more and offer some tentative observations.

Modal-unbiasedness

Wasan (1965 abstract; 1970, pp.120-125) has introduced the notion of a modal-unbiased estimator, that is an estimator whose distribution attains its mode at the true parameter value. The criterion is rather unattractive, in our opinion, partly because the mode of a distribution lacks the intuitive appeal of a median or a mean, partly because its consideration would appear to involve more mathematical complications than the mean (though it might prove to be generally more tractable than the median).

Briefly, an estimator is modal unbiased if the mode of its distribution is the true parameter value. The criterion is related to the loss function:

$$loss\ (\hat{\theta};\ \theta) = \langle\ \hat{\theta} = \theta\ \rangle\ ,$$

but only through a limiting argument starting with a discrete parameter space.

Despite the definition of the MLE as the mode of the likelihood function, *the MLE is not modal-unbiased in general:* the mode of the MLE of the variance of an independently and identically distributed Gaussian variate with known mean, for instance, occurs not at the value of the true variance

σ^2 but rather at $\frac{n-2}{n}\sigma^2$.

Specificity Induced by the Sensitivity Criterion

We will note later on that Barnard and Godambe's sensitivity criterion requires that the mean of the estimating equation equal zero for all parameter values: this certainly serves to narrow the class of possible estimators and would also appear to be properly a specificity criterion, although its effect on the estimator is rather difficult to visualize.

Remarks

Independently and identically distributed observations were assumed in deriving the results on Fisher consistency; however, this does not seem to deny the applicability of the result to wider classes of models. The data could always be regarded as a sample of size one from the appropriate parent distribution, and the MLE could be said to be Fisher-consistent in this case as well, since the sample size is not relevant to the concept.

We note, finally, that the various (competing) specificity criteria will sometimes lead to quite different estimators for the same problem. A remark by Norden (1972, p.342) based on the work of Fend, will serve to illustrate this general fact; for the family

$$p(x;\ \theta) = \theta^{-1/k} \exp\left(-x\theta^{-1/k}\right) \quad (\ x > 0\ )\ ,$$

with k a known integer, the MLE from a sample of size one is x^k whereas an unbiased estimator (one whose variance attains the k-th Bhattacharyya bound) is $x^k/k!$, yet both estimators are specific for θ.

SECTION 2.2 CLOSENESS

In this section we will consider criteria which measure the closeness of the estimator to the parameter. With the possible exception of information-theoretic criteria, it would appear that no optimality results are available in the absence of some specificity criterion; such a criterion is needed in order to rule out "extremely partial" estimators such as the one considered at the beginning of this chapter. It might be argued that specificity and closeness criteria should be studied in pairs, or that, in speaking of a given closeness criterion, one should restrict attention to estimators satisfying a given specificity criterion. However, for simplicity's sake in this section we will consider closeness criteria independently of specificity criteria. This follows an established trend in the literature: for instance, the mean square error (MSE) of a biased estimator is often compared to the MSE (or equivalently the variance) of an unbiased estimator, often without any reference to a specificity criterion being satisfied by the biased estimator. Indeed, there is something to be said for the view that specificity criteria should merely act to screen out undesirable estimators, and that when comparing various estimators one should merely ensure that each estimator satisfies some specificity criterion, that it has some reason for it to be called an estimator.

2.2.1 VARIOUS CONCENTRATION CRITERIA

Although the most common measures of closeness of a parameter to its target value are in terms of a risk function, this is not the most

natural one. A more natural approach would be to select, among several competing estimators, the estimator whose distribution is most concentrated about the true parameter value. We are thus led to consider criteria which are defined strictly in terms of quantities such as $\| \hat{\theta} - \theta \|$. (The Euclidean norm is usually used in this context although there are other possibilities.) When the distribution of an estimator $\hat{\theta}$ is more concentrated about θ than that of an alternative estimator $\tilde{\theta}$, $\hat{\theta}$ is said to be better than $\tilde{\theta}$. In this subsection we will consider three criteria of concentration, and estimators will be called 'better' than others with respect to that measure of concentration.

The nomenclature of concentration criteria appears to be rather fuzzy and the nonce-terms 'strict concentration' and 'pairwise closeness' will be used in our discussion. Finally, we note that the concentration criteria studied here appear to be incompatible with the requirement that an estimation method be invariant under all forms of reparametrization. *As such no optimality properties for the MLE are to be expected here.*

Strict Concentration

The ideal closeness criterion would appear to be what might be called strict concentration: $\hat{\theta}$ is better than $\tilde{\theta}$ whenever, for all $\delta > 0$ and all $\theta \in \Theta$,

$$P[\, \| \hat{\theta} - \theta \| < \delta \,] \geq P[\, \| \tilde{\theta} - \theta \| < \delta \,] \, .$$

However, the criterion is so strong that in most models it seems to rule out all estimators, even when restrictions are placed on the class of alternative estimators $\tilde{\theta}$. The only result we have seen of this type is due to Pitman (1939) for the usual "Pitman estimator" in the special

case of a location model with symmetric likelihood function in all samples. However, Pitman's result relates to fiducial probability, and his result does not appear to have an equivalent frequency interpretation.

Proportional Closeness

The strict closeness criterion has a weak variant in the criterion of proportional closeness: $\hat{\theta}$ is better than $\tilde{\theta}$ for a fixed proportional error δ_o when, for all $\theta \in \Theta$,

$$P[\, \| \hat{\theta} - \theta \| < \delta_o \| \theta \| \,] \geq P[\, \| \tilde{\theta} - \theta \| < \delta_o \| \theta \| \,] .$$

This can be regarded as the minimum risk estimator with loss function:

$$loss\ (\hat{\theta};\ \theta) = \langle\, \| \hat{\theta} - \theta \| < \delta_o \| \theta \| \,\rangle .$$

The criterion has been investigated by Zacks (1966, 1967) and Zacks & Even (1966) but from the negative angle of showing the non-existence of an overall "proportionally closest" estimator. Even if results were available, the criterion is unappealing because the constant δ_o is arbitrary and there does not appear to be any compelling reason for chosing one value of δ_o over another value.

Pairwise Closeness

Another criterion which appears to be derived from the strict closeness criterion is pairwise closeness: in the pair of estimators $(\hat{\theta}, \tilde{\theta})$, $\hat{\theta}$ is better than $\tilde{\theta}$ for a given proportion $\gamma \geq \frac{1}{2}$ (usually, $\gamma = \frac{1}{2}$) if, for all $\theta \in \Theta$,

$$P[\, \| \hat{\theta} - \theta \| < \| \tilde{\theta} - \theta \| \,] \geq \gamma . \qquad (1)$$

The obvious frequency interpretation of (1) is that in at least γ of the samples, the 'better' estimator is closer to the parameter than its competitor. Again, Pitman (1939) proves such a result for fiducial probability. With the fiducial (but not necessarily the frequency) interpretation, it shows that the MLE for a location parameter is inferior to the Pitman 'median' estimator, except where the two coincide. A recent use of the criterion from the frequency standpoint is in Efron (1975) where it is noted that a Stein-type estimator (such as the one defined by equation (1) of the last section) is better than the MLE for a Gauss $(\underset{\sim}{\mu}, \sigma_o^2 I)$ variate.

Admittedly, pairwise closeness is an appealing criterion, but it may be useful to recall the comment by Savage (1972, pp.226, 245): that the main attraction of the criterion would appear to be the conjecture that it should be equivalent in some way to strict closeness; however, a counter-example by Savage disproves this.

More disturbing still is Birnbaum's objection (1961) that the criterion cannot provide a meaningful partial ordering of the class of possible estimators because it depends on the joint distribution of pairs of estimators. It is not excluded, therefore, that one could find some model where:

$$A <^{*} B \ , \quad B <^{*} C \quad \text{and} \quad C <^{*} A \tag{2}$$

(where $A <^{*} B$, e.g., denotes the fact that the estimator A is better than B in the sense of pairwise closeness).

The status of pairwise closeness as a meaningful criterion must, therefore, remain in doubt until (2) is shown to be impossible, i.e., until transitivity is established.

2.2.2 CLOSENESS IN MINIMUM RISK

We turn now to criteria based on loss functions. As noted earlier, the quadratic loss is by far the most popular closeness criterion, both in the asymptotic case and in the finite-sample case; the risk associated with this loss structure is the so-called MSE, and equals the variance plus the square of the bias. It may be recalled that under appropriate regularity conditions, the Cramér-Rao bound provides a useful lower bound on the variance of an unbiased estimator.

As discussed in Section 2.1, in a regular exponential family there would appear to be only one parametrization for which the MLE is unbiased, and under that parametrization the Cramér-Rao bound is attained. It is perhaps fortunate in many exponential families commonly encountered, the mean-value parametrization for which the Cramér-Rao bound is attained is a standard one, whose use is 'natural'. Even when the full standard parameter does not coincide with the full mean-value parameter, it will sometimes happen that the two parameters have in common a subparameter which is of primary interest. Such is the case of the univariate Gaussian distribution under the standard parametrization by the mean and variance, where the MLE for the mean is unbiased and has minimum variance, although the MLE for the variance subparameter is not unbiased and does not have minimum mean square error.

Occasional optimal situations such as the above should not distract us from the fact that *in general the MLE has no optimality property under the MSE criterion* for a parametrization of interest. There are numerous studies in the literature where the performance of the MLE is assessed against that of other estimators, for particular parametrizations of specific

models. It would be outside the scope of this work to delve into particular cases, although given the importance which is usually accorded the criterion, we offer that it might be desirable to have a survey or annotated bibliography of such studies; a good start has been made in that direction in the four-volume monograph by Johnson & Kotz (1969, 1970a, 1970b and 1972), under the topic "Estimation".

Similar studies involving the absolute value criterion are less common but the same comment may be made, that *there is in general no optimality result for the MLE under the minimum-absolute-value criterion* because of the invariance of the MLE under reparametrizations.

2.2.3 SENSITIVITY

An elegant theory of estimating equations and pivotal quantities has been developed by Barnard, Godambe, and others which, for the method of maximum likelihood and under regularity conditions, leads to an optimality property which is valid under a wide class of parametrizations.

The basic idea, due independently to Barnard and to Godambe (see Godambe, 1960, acknowledgement), is to consider estimation methods which can be reduced to solving an equation (or a set of equations, when the parameter is multi-dimensional) of the form $g(x; \theta) = 0$ with the specification that $E_\theta g(x; \theta) = 0$; regularity conditions, both on the quantity q and on the densities p, will be considered for the general situation where θ is multi-dimensional. We shall refer to a quantity g satisfying some specification and regularity conditions as an "unbiased estimating equation". As noted by Godambe (1960), a good unbiased estimating equation should have small variance while at the same time providing

good discrimination between neighbouring parameter values; the latter requirement may be rendered by having $E_\theta \frac{dg(x,\theta)}{d\theta}$ be as large as possible in absolute value. We are, therefore, motivated to consider the variance of the 'standardized estimating equation' $g\left[E_\theta \frac{dg(x,\theta)}{d\theta}\right]^{-1}$ so that we have:

Definition

An unbiased estimating equation g is more sensitive than an alternative unbiased estimating equation h if

$$\frac{E_\theta g^2}{E_\theta\left(\frac{dg}{d\theta}\right)^2} \leq \frac{E_\theta h^2}{E_\theta\left(\frac{dh}{d\theta}\right)^2}$$

holds for all $\theta \in \Theta$.

Result

Under the regularity conditions specified later on in this subsection, the equation for the MLE (i.e., $g = \frac{d \log p(x;\theta)}{d\theta}$) is the most sensitive of all unbiased estimating equations; the MLE equation is essentially unique in being most sensitive: any equation which has the same sensitivity for all values of θ is of the form

$$g(x;\theta) = a(\theta)\frac{d \log p(x;\theta)}{d\theta}.$$

An interesting property shown by Bhapkar (1972) is that the sensitivity of the equation based on a sufficient statistic can never be smaller than the sensitivity of the original equation. In other words, compressing the data to a sufficient statistic may well enhance (but will never

diminish) the performance of the estimating equation; in the case of the likelihood equations no improvement can be had, of course, due to the minimal sufficiency of the likelihood function under regularity.

Regularity Conditions

We consider now the regularity conditions proposed by Bhapkar (1972) which lead to multi-dimensional analogues of the likelihood equation optimality.

With the parameter space Θ an open interval in k-dimensional Euclidean space and an estimator $\hat{\theta}$ such that $g(x; \hat{\theta}(x)) = 0$, assume:

(Conditions on the family of densities)

for all $\theta \in \Theta$

A.1) $\nabla_\theta \log p(x; \theta)$ and $\nabla_\theta \nabla_\theta^T \log p(x; \theta)$ exist a.e. $\lambda(x)$;

A.2) both $\int p(x; \theta)\, d\lambda(x)$ and $\int \nabla_\theta \log p(x; \theta)\, d\lambda(x)$ can be differentiated with respect to θ under the integral sign;

A.3) $E_\theta \left[(\nabla_\theta \log p)(\nabla_\theta \log p)^T \right]$ is positive definite.

(Conditions on the estimating equation)

for all $\theta \in \Theta$

B.1) $E_\theta g(x; \theta) = 0$;

B.2) $\nabla_\theta g(x; \theta)$ exists a.e. $\lambda(x)$;

B.3) $\int g(x; \theta)\, p(x; \theta)\, d\lambda(x)$ is differentiable under the integral sign;

B.4) $E_\theta \left[(\nabla_\theta g(x; \theta))(\nabla_\theta g(x; \theta))^T \right]$ is positive definite;

B.5) $E_\theta [\, g(x; \theta)(g(x; \theta))^T]$ exists finitely.

Multi-dimensional Case

The multi-dimensional analogue of (2) uses the matrix

$$J_g(\theta) = [E_\theta \nabla_\theta g]^{-1} V_\theta(g(x; \theta)) [(E_\theta \nabla_\theta g)^T]^{-1} \quad (4)$$

(our notation deviates from Bhapkar's at this point).

Result

With g^* *denoting the set of likelihood estimation equations*

$$g^*(x; \theta) = \nabla_\theta \log p(x; \theta) ,$$

the matrix

$$J_g(\theta) - J_{g^*}(\theta) \quad (5)$$

is at least positive semi-definite for any other unbiased estimating equation g.

Bhapkar proposes two scalar analogues of (3) based on the characteristic roots of (5). Thus we are led to two subsidiary optimality properties for g^*.

$$det\ J_{g^*}(\theta) \leq det\ J_g(\theta) \quad (6)$$

and

$$trace\ J_{g^*}(\theta) \leq trace\ J_g(\theta) .$$

Equality, in (6) at least, is ruled out except for g equal to a multiple of the likelihood estimating equation.

We note in passing that there exists a multi-dimensional analogue

of enhanced performance under compression to a sufficient statistic.

Discussion

In our view there are three main limitations to the usefulness of the results concerning sensitivity.

The first is that the regularity conditions (A.1, A.2, A.3) exclude models where the density has a carrier depending on the parameter. Indeed, in such models the MLE, if it exists meaningfully, is not in general a solution of the likelihood equations so that not much is to be expected from any extensions of the criterion in this direction.

Second, condition (B.1) serves to rule out wide classes of estimation methods. If the estimating equation $g(x; \theta) = 0$ produces an estimator $\hat{\theta}$ with bias $E_\theta g(x; \theta) = h(\theta)$, then the unbiased version of that estimating equation is $g^*(x; \theta) = g(x; \theta) - h(\theta)$ which (in the nontrivial case where the bias $h(\theta)$ is not constant) should lead to an estimator $\tilde{\theta}$ different from $\hat{\theta}$. Alternatively, one could retain the biased estimating equation but modify formulas (3) and (5). For a one-dimensional parameter simple algebra leads to the inequality:

$$\frac{E_\theta(g - Eg)^2}{\left[E_\theta\left(\frac{dg}{d\theta}\right) - \frac{d}{d\theta} E_\theta g\right]^2} \geq \frac{E_\theta\left(\frac{d \log p}{d\theta}\right)^2}{\left[E_\theta \frac{d \log p}{d\theta}\right]^2} \quad . \tag{7}$$

This resembles the version of the Cramér-Rao inequality for biased estimators - of which it is a generalization. We fancy that the result then would lose its appeal to many for reasons similar to those which make the 'biased' version of the Cramér-Rao inequality unattractive: not only does the inequality fail to address itself to departures from the target

value $g = 0$, but also the impact of the derivative of the bias is difficult to assess.

Lastly, the optimality criterion considered here differs from other criteria we are considering in that it properly concerns the method of estimation rather than the estimator. Even after gaining some familiarity with inequalities such as (3), the author finds himself wondering "Yes, the MLE comes from the most specific estimating equation in its class: but what does that tell me about the MLE itself?"

Against these reservations (and the last one may be due simply to insufficient familiarity with the present concepts), the criterion has the great advantage (as noted by Barnard, 1973) of being invariant to reparametrizations, subject, of course, to the change of parameters being smooth enough to respect the regularity conditions.

Furthermore, the criterion bears an intimate relation to pivotal quantities (to the extent that we could have replaced 'estimating equations' by 'pivotal quantities' in the above discussion; we did not do so in order to underline the fact that the criterion does not refer to estimators as such but to a method of obtaining estimators). Pivotal quantities provide a sound approach to inference since, as noted by Kempthorne & Folks (1971, p.338) among others, stable pivotals - those whose distribution is altogether independent of the parameter - seem to provide the only entry into exact distribution theory, and since much of the approximate theory is likewise based on approximately stable pivotals.

2.2.4 INFORMATION-THEORETIC CLOSENESS

We approach the topic of this subsection with some diffidence because the approach taken by many authors on the subject is rather wider than that of parametric estimation: (see, for instance, Good, 1963; Dutta, 1966; Kullback, 1968, pp.37-39; and Gokhale & Kullback, 1978, p.17). These authors, and others, use information-theoretic arguments to resolve the global problem of specifying the model, where the specification in many instances is so deep that it includes the actual setting of values for the model's parameter; this last operation is, of course, our own problem of point estimation. Inconsistencies, or at least meaningless results, are to be feared when the concepts appropriate to the wider problem are constricted to the narrower one. Nevertheless, it may be worthwhile to sketch, at least, some points of contact between the information-theoretic approach and maximum-likelihood estimation.

We begin by offering some standard definitions from Information Theory. In general, Information Theory may be said to be primarily concerned with discrepancies between specified probability distributions. Let F and G be two probability distributions, corresponding to probability measures μ_F and μ_G, where μ_F is absolutely continuous with respect to μ_G and where both are absolutely continuous with respect to a σ-finite dominating measure λ. A measure of the discrepancy between F and G as seen from the distribution F is the Kullback-Leibler Separator already encountered in Subsection 2.1.4:

$$I(F : G) = \int \log \left[\frac{dF}{d\lambda}(x) \left(\frac{dG}{d\lambda}(x) \right)^{-1} \right] dF(x) .$$

Other names for $I(F : G)$ are 'the mean information for discriminating in favor of F against G' (Kullback, 1968, p.5) or 'the information of G relative to F'. We note in passing that I is not a symmetric operator: $I(F : G) \neq I(G : F)$ for most F and G; and that $I(F : G) \geq 0$. A motivation for considering I is given in Savage (1972, pp.48-49) in that I is the expected value of the logarithm of the likelihood-ratio statistic for two competing hypotheses F and G. The theory of $I(F : G)$ is developed extensively by Kullback (1968).

A related concept is that of the entropy of a probability distribution F:

$$H(F) = - \int \log \frac{dF}{d\lambda}(x)\, dF(x) .$$

The motivation for H as a measure of the disorder in the distribution F has been given by Shannon (1948, p.392). Analogous motivations are reviewed by Rényi (1961). In the remainder of this subsection, the arguments of I and H will be cumulative distribution functions or their densities as convenient in the context.

Discrete Distributions

We preface this part of the discussion with the remark that the discussions in the literature are in the context of contingency tables and do not necessarily extend to models with an infinite number of cells.

In discrete models with n_x observations of the value x and $n = \Sigma_x n_x$ observations in all the 'observed distribution', $q(\cdot)$ which takes the value $q(x) = n_x/n$ for the cell x, is absolutely continuous with respect to all the possible model distributions $p(\cdot; \theta)$

which are of interest. The exception is those values of θ for which $p(x; \theta) = 0$ in a cell x in which a nonzero count was observed $n_x > 0$; in such cases the data provide a crucial test against such values of θ (the hypothesis indexed by θ being incompatible with the data). It seems reasonable that we would want to exclude such values of θ in any estimation process. It therefore makes sense to speak, at least formally, of the information of $p(\cdot; \theta)$ in favour of the data:

$$I(q : p(\cdot; \theta)) = \sum_x \frac{n_x}{n} \log \frac{n_x/n}{p(x; \theta)} . \qquad (8)$$

It is easy to see from (8) that for fixed $q(\cdot)$, the value of θ for which $I(q ; p(\cdot; \theta))$ is minimized, is precisely the value which maximizes the logarithm of the likelihood function, so that we have the formal interpretation that the MLE selects the distribution $p(\cdot; \theta)$ so as to minimize the discrimination information in the sample relative to the class of possible distributions (see Bishop, Fienberg & Holland, 1975, p.346, for example). More loosely, the estimated distribution $p(\cdot; \theta)$ is 'closest to the data' in some sense.

Against this it should be noted that the other variant $I(p(\cdot; \theta) : q)$ is minimized by the 'minimum discrimination information estimate' (in the nomenclature of Gokhale & Kullback, 1978) or by what Bishop, Fienberg & Holland (1975) term the 'modified minimum discrimination information estimate' (MMDIE). In general the MLE is not equal to the MMDIE so that, using this second 'optimality criterion', the MLE is not in general optimal. The class of situations in which the two estimates are equivalent is called the 'internal constraint problem'. It corresponds to

those situations where the data happen to fit the model perfectly, the observed counts satisfying all the constraints imposed by the model.

Exponential Models

Exponential families are the only other instance where we have found what we believe to be valid results. In the information-theoretic approach exponential distributions are distinguished as that class of probability distributions $p(\cdot;\ \theta)$ (denoted here by the densities) which assigns finite value to the expectation of the given statistic t and for which $I(\ p(\cdot;\ \theta) : q\ )$ is minimized for a fixed 'reference' distribution $r(\cdot)$. That is:

With $G_\theta = \{\ p(\cdot) : \int t(x)p(x)\ d\lambda(x) = \theta\ \}$,

G_θ being a class of probability densities with respect to λ, let $p(\cdot;\ \theta)$ be the element of G_θ which minimizes $I(p^*;\ r)$ for $p^* \in G_\theta$; then:

$$p(x:\ \theta) = e^{\theta^T t(x)}\ c(\theta)r(x).$$

It is shown by Kullback (1968, p.94) that the MLE for a given sample selects θ so as to minimize $I(\ p(\cdot;\ \theta) : r\ )$. Another result along those lines is given by Simon (1973). With a given sample yielding an (unconstrained) MLE $\hat{\theta}$ when the parameter space is Θ, let Θ_1 be a subset of Θ. Then the MLE $\hat{\theta}_1$ for θ restricted to Θ, selects that member of $\{\ p(\cdot;\ \theta) : \theta \in \Theta_1\ \}$ which minimizes $I(\ p(\cdot;\ \theta) : p(\cdot;\ \hat{\theta})\ )$ over $\theta \in \Theta_1$. Thus the constrained MLE $\hat{\theta}_1$ is 'closest' to the unconstrained MLE $\hat{\theta}$, just as the latter is 'closest' to the reference distribution $r(\cdot)$.

Discussion

Another, rather tantalizing result in information-theoretic closeness has been presented by Kriz & Talacko (1968). However, there are serious problems with the proof given, and we have prefered to discuss the question in Appendix C.

The remaining results in this subsection are interesting, but it is difficult to translate them into a claim that the MLE is close to the parameter in any reasonable sense. In particular, we find that, in the discrete case, averaging the log-likelihood over a legitimate distribution, $I(\, p(\cdot;\, \theta) : q \,)$, is more meaningful than averaging over the 'observed' distribution, so that we would prefer the optimality property satisfied by the MMDIE to the one satisfied by the MLE.

CHAPTER 3 SUFFICIENCY AND INVARIANCE

In this chapter sufficiency and various types of invariance are considered. The reason for dealing with both properties together will become clear in the second section, where sufficiency plays a subsidiary role.

SECTION 3.1 SUFFICIENCY

3.1.1 INTRODUCTION

The concept of sufficiency was introduced into statistics by Fisher (1920), although the term itself was essentially introduced in a later article (1922). Some aspects of the 'prehistory', so to speak, of sufficiency are considered by Stigler (1973 and 1976). It would appear that a parallel concept in Statistical Mechanics emerged earlier (see Mandelbrot, 1962), however, the context there is much wider and similar to the situation discussed at the beginning of subsection 2.2.4.

Briefly, a statistic t is sufficient for the parameter θ if for any other statistic s, the distribution of s conditional on $t = t_o$ is the same for all values of θ. A sufficient statistic t is minimal sufficient if it can be written as a function of any sufficient statistic.

There is some question as to whether sufficiency has any relevance in the context of point estimation (Savage, 1976, p.459). An estimator

is principally a function which points to one member of $\{p(\cdot; \theta) : \theta \in \Theta\}$ as being a reasonable approximation to the underlying distribution which gave rise to the data. However, inasmuch as an estimate will often have to be used alone, as it were, in lieu of the full data set (as in the examples cited in Chapter 1), the estimator is also a summary of the data (Rao, 1962a) and sufficiency is relevant at this point since any information which is lost in this condensation is irrelevant to the model. Indeed, sufficiency was introduced in the specific context of maximum-likelihood estimation.

Fisher appears to have believed (1925) that when a sufficient statistic exists, the MLE must be sufficient. That this is not so can be seen from an example of Barndorff-Nielsen to be discussed later, as well as from an example of Savage (1976, pp.460-461). The latter example is significant in that it also shows (at least in some finite, discrete parametric families) that it is possible to construct a Fisher-consistent, sufficient estimator in a situation where the MLE itself is not sufficient.

The theory of the sufficiency of MLEs, as indeed the theory of sufficiency as a whole, seems to be rather incomplete at this moment, although we may have missed some important developments in the literature. The remainder of this subsection will consider what can be said about sufficiency in general models.

The next three subsections will consider three classes of models with identically and independently distributed observations. We hesitate somewhat in presenting this material, since what is correct in those subsections appears in two recent monographs (Barndorff-Nielsen, 1978; Huzurbazar,

1976), while the remainder is hardly more than conjectures supported by heuristic arguments. Nevertheless, it may be useful for the sake of completeness to abstract the results from the above monographs and to complement them with some (admittedly shaky) additional results. Throughout this section, p will denote the density of the data and q, the density of one observation (i.e., the density of the parent distribution).

An important criterion for the existence of a sufficient statistic is the factorization criterion (Fisher, 1922; see Bahadur, 1954 for a fuller proof). The criterion applies to all models (not necessarily with identically and independently distributed observations) where the family $\{P(\cdot|\theta);\ \theta \in \Theta\}$ of distributions is dominated by a σ-finite measure λ. It states that a sufficient statistic $t(x)$ exists if and only if there exist non-negative functions h on the range of x and g on the range of t with the compound function $g(t(\cdot);\ \theta)$ measurable as a function of x, and h functionally independent of θ, so that:

$$\frac{dP(\cdot|\theta)}{d\lambda}(x) = p(x;\ \theta) = g(t(x);\ \theta)h(x). \qquad (1)$$

The proof of the criterion (e.g., Lehmann, 1959, pp.48-50) indicates that when t is sufficient, the functions g and h may be taken to be densities.

Although we are not concerned here with models where $\{P(\cdot|\theta) : \theta \in \Theta\}$ is not dominated, it may be noted that minimal sufficiency is a rather uninteresting property in such models. Burkholder (1961) has shown that in such cases, there may exist statistics s and t, with t being a compression of s, $t = f(s)$, such that t is sufficient but s is not: in other words, compressing the data may add information.

From (1) it is easy to derive the result that the MLE must be a function of any sufficient statistic. In the context of sufficiency, however, this does not say very much, since the function is not necessarily one-to-one, so that sufficiency may be lost. The result is discussed at greater length under the topic of invariance to data transformation (subsection 3.2.1).

3.1.2 CONSTANT CARRIER, CONTINUOUS CASE

The theory of sufficiency seems to be best established for parent distributions which are absolutely continuous with respect to Lebesgue measure and whose densities have a carrier set which is the same for all members of the parametric family. In this setting, and with regularity conditions, it can be shown (Koopman, 1936; Pitman, 1936; also: Fisher, 1934; Darmois, 1935) that the only parent distributions which admit a sufficient statistic of constant dimensionality for all sample sizes are of the exponential form with densities:

$$g(x;\ \theta) = \exp[\ t(x)^T \phi(\theta)\ -\ \kappa(\theta)\ +\ g(x)\ ]\ . \qquad (2)$$

It is convenient to require that the components of t be affinely independent:

$$\Sigma\ a_i t_i(x) = a_o \ \text{a.e.}\ \lambda(x) \Rightarrow a_i \equiv 0$$

and that $\phi(\cdot)$ range over a k-dimensional subset of Euclidean space which contains an open (k-dimensional) interval, $k = dim\ \phi = dim\ t$. It is easily seen that, when $dim\ \phi = dim\ \theta$, $t(x)$ is sufficient for θ, since the conditional density of x given $t = t_o$ is

$$\int_{t(x)=t_o} g(x;\,\theta)dx = \int_{t(x)=t_o} \exp[\,g(x)\,]dx\ .$$

It is also true that t is minimal sufficient.

When *dim* $\phi >$ *dim* θ, the exponential family (2) may be said to be curved, in the nomenclature of Efron (1975). In curved exponential families it appears to be a general rule that the MLE is not sufficient, although a completely rigorous proof eludes us. With T = interior of the range of t, assume $P(T|\theta) > 0$ for all $\theta \in \Theta$. Then with probability greater than zero, the MLE $\hat{\theta}$ is a solution of the likelihood equation:

$$\nabla \log p(x;\,\theta) = \sum_{i=1}^{n} \nabla \log q(x_i;\,\theta)$$

$$= n\nabla\,\phi(\theta)^T\bar{t} - n\nabla\,\kappa(\theta) = 0,$$

where $\bar{t} = \Sigma t(x_i)/n$. Thus:

$$\nabla\,\phi(\hat{\theta})^T\bar{t} = \nabla\,\kappa(\hat{\theta}).$$

Now $\nabla\phi^T$, a matrix of dimension *dim* $\theta \times$ *dim* ϕ, has rank no larger than *dim* θ, and therefore rank strictly smaller than *dim* $\bar{t}$. Hence the MLE will be a strict contraction of the minimal sufficient statistic $\bar{t}$, and cannot itself be sufficient. Even when *dim* ϕ = *dim* θ, however, it need not be true that the MLE will be sufficient. A trivial case must first be considered; when the parameter space Θ is smaller than it could be, it will often happen that the restricted MLE will occur on the boundary of Θ and that it will be a many-to-one function of $\bar{t}$. For example, when the parent distribution is univariate Gaussian with unknown mean μ and known variance σ_o^2, and when the parameter space is restricted to

$\{\mu : \mu \geq 0\}$, then with probability $\Phi(\mu) > 0$, the sample mean $\bar{x}$ will be negative and the MLE takes value zero, so that the MLE is not a one-to-one function of the minimal sufficient statistic $\bar{x}$. Even making Θ as large as possible does not ensure that the MLE will be sufficient. Consider an example due to Barndorff-Nielsen (1978, pp.152-153). The parent distribution is

$$q(x;\ \theta) = kx^{-k-1}\ e^{\theta x - \kappa(\theta)} \quad (\ x > 1\ ).$$

where k is a known constant greater than one. For general θ, the function $\kappa(\theta)$ cannot be represented in closed form in terms of elementary functions, although when $\theta \leq 0$, κ can be properly defined as

$$\kappa(\theta) = \log\left[\int_1^\infty kx^{-k-1}\ x^{\theta x}\ dx\right] . \qquad (3)$$

Furthermore, when $\theta > 0$ the integral in (3) diverges, so that, at its fullest extent, $\Theta = (-\infty,\ 0]$.

In this model, the sample mean is a minimal sufficient statistic, but whenever the sample mean is greater than $k/(k-1)$, the corresponding MLE is equal to zero. To see this, note that:

$$\frac{d\kappa}{d\theta} = \int_1^\infty kx^{-k}\ e^{\theta x}\ dx e^{-\kappa(\theta)} = E_\theta x .$$

(This is true in general of exponential families.) Also,

$$\begin{aligned}\frac{d^2\kappa}{d\theta^2} &= \frac{d}{d\theta}\int kx^{-k}\ e^{\theta x - \kappa(\theta)}\ dx \\ &= \int kx^{-k+1}\ e^{\theta x - \kappa(\theta)}\ dx - \int kx^{-k}\ e^{\theta x - \kappa(\theta)}\ dx \cdot \frac{d\kappa}{d\theta} \\ &= E_\theta x^2 - E_\theta x\ E_\theta x\end{aligned}$$

and so

$$\frac{d^2\kappa}{d\theta^2} = V_\theta(x) > 0 .$$

This means that $E_\theta x$ is an increasing function of θ, so that its greatest value is

$$\max\{ E_\theta x : \theta \in \Theta \} = E_0 x = \int_1^\infty k x^{-k} \, dx = k/(k-1) .$$

Now if $\bar{x} > k/(k-1)$, the likelihood function is strictly increasing, since

$$\frac{d}{d\theta} \log \prod_{i=1}^n q(x_i; \theta) = n \frac{d}{d\theta} (\theta\bar{x} - \kappa(\theta) - \text{constant})$$

$$= n (\bar{x} - \frac{d\kappa}{d\theta} (\hat{\theta})) > 0 .$$

Therefore, the MLE fails to be sufficient, even though a sufficient statistic exists. This occurs because the model allows realizations of the sufficient statistic t which cannot be values of $E_\theta t$ for any θ.

A statement of conditions for the MLE to be sufficient requires some nomenclature adapted from Barndorff-Nielsen (1978).

- When $\phi(\theta)$ is a one-to-one mapping of θ, it is possible to reparametrize the family so as to have

  $$q(x; \theta) = \exp\{ t(x)^T\theta - \kappa; (\theta) \} b(x) ;$$

  when this can be done the parameter θ is called the <u>natural</u> parameter and its corresponding parameter space Θ, the <u>natural parameter space</u>.

- Θ is said to be <u>full</u> when

  $$\Theta = \{ \theta : \int \exp(t(x)^T\theta) b(x) dx < \infty \} ,$$

  i.e., when the parameter space is as large as possible.

- The finite boundary of the full natural parameter space is the set of finite θ, every neighbourhood of which contains points in both Θ and in its complement.

- κ is steep if for any sequence $\{\theta_i\}_{i=1}^{\infty}$ converging to a point on the finite boundary of Θ, $\lim_{i\to\infty} \|\nabla\kappa(\theta_i)\| = \infty$. (Differentiability of κ is guaranteed by Theorem 9 in Lehmann, 1959, p.52.) When the finite boundary is void (i.e., when the natural parameter space is the Euclidean space), κ is steep by definition.

Also, denote:

S = convex hull of the range of $t(\cdot)$,

C = *closure* S

B = finite boundary of S.

Result (Barndorff-Nielsen, 1978, p.152)

When $P(B|\theta) = 0$ *and* Θ *is full, the MLE is a sufficient statistic if and only if* κ *is steep.*

An equivalent condition for steepness (and hence for sufficiency of the MLE) is $P(B|\theta) = 0$ and

$$\tau(\textit{interior}\ \Theta) = \textit{interior}\ C,$$

where $\tau(\theta)$ is the mean-value parametrization $\tau(\theta) = E_\theta t$. (Barndorff-Nielsen, 1978, p.142.) It may be noted that the above results are valid whatever the dominating measure, so long as the carrier of the exponential family distributions is constant.

3.1.3 VARIABLE CARRIER, CONTINUOUS CASE

The regularity conditions used to derive the standard characterization of distributions admitting sufficient statistics exclude families of continuous distributions whose carrier set depends on the parameter. The general case, where the carrier set is arbitrary, seems to be very complicated (see, e.g., Fraser, 1963). However, for the important situation where the observations are univariate and where the carrier set is an interval (possibly a half-line), Huzurbazar (1976, pp.95-187) has developed a rather complete theory to complement the Koopman and Pitman results. There is, however, no discussion of the sufficiency of the MLE in Huzurbazar's treatment.

Huzurbazar's results may be condensed to two types of model. Let the carrier set be denoted $(a(\theta), b(\theta))$. In the first type, a and b depend on θ through a one-dimensional subparameter α, so that the carrier set may be denoted $(a(\alpha), b(\alpha))$. When a and b are differentiable monotone functions in α (such that as α increases, a is non-increasing and b is non-decreasing or vice-versa), Huzurbazar's theory shows that the only (parent) parametric families which admit a sufficient statistic have density of the form:

$$q(x;\theta) = \exp[\, t(x)^T \phi(\theta) - \kappa(\theta) + g(x) \,] \quad (a(x) < x < b(\alpha)) \ . \qquad (4)$$

(When θ is itself one-dimensional, the inner product $t^T\phi$ in (4) vanishes.) A sufficient statistic for θ in (4) is then the set $(\tilde{\alpha}, t(x))$ where

$$\tilde{\alpha} = \max(\inf\{ \alpha : a(\alpha) = x_{(1)} \}, \inf\{ \alpha : b(\alpha) = x_{(n)} \}) \ ,$$

when a is non-increasing and b is non-decreasing. (When the directions

of a and b are reversed, the theory can be applied by reparametrizing $\alpha \to -\alpha$.)

The question of the sufficiency of the MLE θ in (4) does not appear to have been considered in general, and we only sketch a proof. Let us assume a very simplified version of (4), where it is possible to reparametrize the model $\theta \to (\alpha, \phi)$ so that:

$$q(x; \alpha, \phi) = \exp[\, t(x)^T\phi - \kappa(\alpha, \phi) + g(x)\,] \quad \langle\, a(\alpha) < x < b(\alpha) \,\rangle \,. \tag{5}$$

Also, assume both a and b are strictly monotone functions.

It would appear that sufficiency results could be obtained by a two-stage argument, first fixing α at α_o and applying the constant-carrier result, and then letting α_o tend to $\tilde{\alpha}$. In order for the argument to apply, the range of ϕ for fixed α would have to be

$$\Phi(\alpha) = \{\, \phi : \int_{a(\alpha)}^{b(\alpha)} \exp[\, t(x)^T\phi + g(x)\,]dx < \infty \,\} \,;$$

it must also be assumed that the finite boundary of $\Phi(\alpha)$ has probability zero. When this is so, the steepness condition, $\left|\frac{\partial \kappa(\alpha, \phi)}{\partial \phi}\right| \to \infty$ on the finite boundary, would guarantee that the MLE $\phi(\alpha_o, x)$ for the subparameter ϕ conditional on $\alpha = \alpha_o$, would be a one-to-one function of t. It can also be established that $\kappa(\alpha, \phi)$ is increasing in α, so that the likelihood function for fixed ϕ is decreasing in α. Therefore, since $\tilde{\alpha}$ is the smallest value of α which is consistent with the data, the likelihood function should attain its absolute maximum at $(\tilde{\alpha}, \hat{\phi}(\tilde{\alpha}, x))$. Then $(\tilde{\alpha}, \hat{\phi})$ would always be a one-to-one function of the sufficient statistic $(\tilde{\alpha}, t)$, and the MLE would be a sufficient statistic.

The second type of parent distribution which admits a sufficient statistic is where the endpoints depend on a two-dimensional subparameter,

say (α, β). Then Huzurbazar's theory shows that the distributions must have density of the form:

$$q(x; \theta) = \exp[\, t(x)^T \phi(\theta) - \kappa(\theta) + g(x) \,] \quad (a(\alpha, \beta) < x < b(\alpha, \beta)) . \qquad (6)$$

(When θ is two-dimensional, the inner product $t^T \phi$ once again vanishes.) Huzurbazar's theory shows that a minimal sufficient statistic for θ is: $(x_{(1)}, x_{(n)}, t(x))$.

It is possible to consider the sufficiency of the MLE in models where (6) can be reparametrized to:

$$q(x; \alpha, \beta, \phi) = \exp[\, t(x)^T \phi - \kappa(\alpha, \beta, \phi) + g(x) \,] \quad (\alpha < x < \beta) , \qquad (7)$$

where it is further assumed that (α, β) ranges over a two-dimensional interval $A \times B$. When the range of ϕ for fixed (α, β) is

$$\Phi(\alpha, \beta) = \{ \phi : \int_{\alpha}^{\beta} \exp[\, t(x)^T \phi + g(x) \,] dx < \infty \} ,$$

and when the finite boundary of $\Phi(\alpha, \beta)$ has probability zero one can use a two-stage argument similar to the above. It can be seen that $\kappa(\cdot, \beta, \phi)$ is strictly decreasing while $\kappa(\alpha, \cdot, \phi)$ is strictly increasing, so that if the maximum of the likelihood for fixed $(\alpha, \beta) = (\alpha_o, \beta_o)$ occurs at $\hat{\phi}(\alpha_o, \beta_o, x)$, then the absolute maximum must occur at the point $(\hat{\alpha}, \hat{\beta}, \hat{\phi}) = \left(x_{(1)}, x_{(n)}, \phi(x_{(1)}, x_{(n)}, x)\right)$. When the steepness condition is satisfied, $\hat{\phi}(\alpha_o, \beta_o, x)$ is a one-to-one function of t, so $(\hat{\alpha}, \hat{\beta}, \hat{\phi})$ is a one-to-one function of the sufficient statistic $(x_{(1)}, x_{(n)}, t)$. Hence under the above conditions, it would appear that the MLE must be sufficient.

3.1.4 DISCRETE CASE

The regularity conditions of the Koopman and Pitman result (2) also exclude models where the observations are discrete. Jeffreys (1960) has offered a proof that discrete parent distributions which admit sufficient statistics are of the exponential type. However, his proof asserts that a set of non-linear equations must have a solution, and this fact is not at all transparent to us. It may, however, be remarked that all discrete distributions with finite carriers admit a sufficient statistic of constant dimension (namely, the vector of counts of observations in each cell).

The theory developed by Barndorff-Nielsen (1978) for the sufficiency of the MLE cannot be applied to the discrete case, except when the carrier set in totally unbounded - e.g., a family of distributions having the set of all integer numbers (positive and negative) as its carrier set. When the sufficient statistic $t(x)$ is restricted, say $t(x) \geq 0$, then because of discreteness there must be at least one mass point (e.g., $t = 0$) on the boundary of S, the convex hull of the range of $t(\cdot)$. Hence, the conditions $P(B|\theta) = 0$ is not satisfied.

Barndorff-Nielsen (1978, p.155) shows that when the range of t is finite, the MLE for the mean-value parameter $\tau = \tau(\theta) = E_\theta t$ in:

$$q(x;\ \theta) = \exp[\ t(x)^T\theta - \kappa(\theta) + g(x)\ ],$$

is $\hat{\tau} = t$, so that it may be claimed that the MLE is sufficient in such cases (the steepness of κ is ensured by the fact that with the range of t being finite, the full natural parameter space must be the Euclidean space of the appropriate dimension). Some caution is necessary with this

result, however, since points τ on the border B may correspond to infinite parameter values in the 'natural' parametrization. In a simple model such as the Bernoulli, the mean-value parametrization is the standard one to use; in other models, however, an MLE $\hat{\tau} \in B$ may not be reasonable, and the full parameter cannot be estimated by maximum-likelihood. (See Barndorff-Nielsen, 1978, pp.156-158, for an example involving logistic regression with intercept α and slope β, where for some samples only $\alpha + \beta$ can be estimated, in other, only α.)

3.1.5 CONCLUSION

The results of this survey of sufficiency properties are rather disappointing. The cases convered are far from exhaustive, and even in those cases, the theory is at places sketchy. Perhaps the best conclusion would be that the MLE is not necessarily sufficient, even in models which admit a sufficient statistic whose dimension is that of the parameter.

SECTION 3.2 INVARIANCE

The term 'invariance' will be used in this section in two distinct meanings:

(1) An estimator which is computed from a transformed version of the data (using the model induced by the transformation) may coincide with the estimator computed from the original data and model, where both estimators are derived according to the same method. In symbols:

$$y = t(x) \Rightarrow \hat{\theta}_y(t(x)) = \hat{\theta}_x(x) \ .$$

(2) When the distribution is reparametrized, the estimator of the transformed parameter may equal the transformation of the estimator of the original parameter, both estimators being derived from the same method. In symbols:

$$\phi = \psi(\theta) \Rightarrow \hat{\phi}(x) = \psi(\hat{\theta}(x)) \ .$$

The third subsection will consider the situation where a conjunction of the transformations in points (1) and (2) is such that the model resulting from the double transformation is the same as the original model.

The basic material in this section is so standard that the historical notes will be omitted. The presentation, particularly in subsection 3.2.1, deviates somewhat from what we have seen in the literature.

3.2.1 INVARIANCE WITH RESPECT TO TRANSFORMATIONS OF THE DATA

Transformations of the data occur routinely in at least two settings. First, the data may be rescaled in some way, perhaps to facilitate computations. One example is taking the logarithm of observations which are

assumed to follow the lognormal distribution. Second, the data may be compressed, as is done regularly with independently and identically distributed discrete data where instead of reporting, say, that the first observation fell in cell number 4, the second fell in cell number 1 and so on, it is standard practice to report only the count of events in each cell (i.e., essentially, the order statistic). Condensation has occurred here since the order in which the events occurred cannot be recovered from the order statistic. Here we will consider a wider class of transformations which includes both of the above types. In general, consider measurable transformations $t : \Xi \to \Upsilon$, where Υ may be Ξ itself. For each t the transformation will carry the variable x with density $p(\cdot; \theta)$ into a random variable $y = t(x)$ with density $q(\cdot; \theta)$. For general t, the distribution in the transformed model $\{ y; q(\cdot; \theta); \theta \in \Theta \}$ will not be identified: a trivial example is $t(x) \equiv$ constant, where $q(\cdot; \theta_1) \equiv q(\cdot; \theta_2)$ for all $\theta_1, \theta_2 \in \Theta$. However, a sufficient condition for $\theta \in \Theta$ to be identifiable in $q(\cdot; \theta)$ is that $t(x)$, considered as a statistic, be sufficient for θ. The proof of this is easily seen from the factorization criterion (formula (1) of Section 3.1). Since:

$$q(t(x); \theta)\, r(x) = p(x; \theta)$$

then $q(\cdot; \theta_1) \equiv q(\cdot; \theta_2)$ if and only if

$$p(\cdot; \theta_1) \equiv p(\cdot; \theta_2)$$

so that θ is identifiable under q if it is identified under the original model p. We note, however, that some non-sufficient transformations may also produce identifiable models.

Definition

An estimation method is said to be invariant under the family T of transformations which preserve sufficiency if for every $t \in T$ and for almost every sample x the method produces estimators $\hat{\theta}_x$, $\hat{\theta}_y$ respectively for the models' $\{ x; p(\cdot; \theta); \theta \in \Theta \}$ and $\{ y; q(\cdot; \theta); \theta \in \Theta \}$ such that $\hat{\theta}_x(x) = \hat{\theta}_y(t(x))$. We have the following simple

Result

The method of maximum likelihood is invariant under all transformations of the data which preserve sufficiency.

Remark

Of the two possible approaches - finding the mode of the likelihood function and finding the root of the likelihood equation which maximizes the likelihood function - the same approach must be taken for both models.

Proof

Using the Neyman factorization criterion, it is seen that the likelihood function for the transformed model, using $t(x)$, must equal the likelihood function for the original model based on x, up to a constant multiplicative factor. Therefore, the maximum-likelihood estimator must be the same in both cases.

The result is hardly new, of course: it is merely a transposition in the setting of invariance of the familiar result that the MLE is a function of every sufficient statistic. A consequence is

Corollary 1

The method of maximum-likelihood will yield the same estimator under all one-to-one measurable transformations of the data.

This last result has been proved under the somewhat more restrictive assumptions of the 'inverse function theorem', that the transformation be differentiable and have a differentiable inverse (e.g., Sverdrup, 1967, p.123). Both the Result and Corollary 1 concern the method of maximum likelihood rather than the estimator it produces. It may, therefor, be worthwhile to state separately a property of the MLE itself.

Corollary 2

When the observations are independently and identically distributed the maximum-likelihood estimator remains invariant under all permutations of the observations.

Corollary 2 is, in a sense, a minimal property one should expect of an estimator in this situation, and indeed we cannot think of any standard estimating procedure which would take the order of the observations into account when independence of identically distributed observations is specified.

3.2.2 INVARIANCE WITH RESPECT TO TRANSFORMATIONS OF THE PARAMETER

To a certain extent, it may be argued that the parametrization chosen in the model being entertained is inessential, serving merely as an indexing scheme for the various probability distributions and that the goal of estimation is to select the distribution which is most appropriate to represent the data that have been obtained. While this is an extreme

view, it is often the case that several parametrizations are possible and meaningful for the same family of distributions. The two-parameter Gaussian distributional family is usually indexed by its mean and variance (μ, σ^2) but (μ, σ) would, in some sense be an even more natural choice, since both of its components are of the same physical dimension and σ is a scale parameter. Other possible parametrizations are (μ, ε) where $\varepsilon = 0.6745\sigma$ is the probable error, and (μ, τ), where $\tau = \sqrt{2}/\sigma$ is the modulus of precision. The exponential distribution with density $p(x; \theta) = \exp(-x/\theta)\ (x > 0)$ might be parametrized by the frequency of failure $\phi = 1/\theta$ as well as by the expected lifetime θ. In a more general vein, for exponential families it is more convenient to work with the 'natural' parametrization even when this differs from the standard parametrization: for the two-parameter Gaussian, the 'natural' parameter corresponding to the usual sufficient statistic $(\Sigma x, \Sigma x^2)$ is

$$\theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) .$$

Generally, when the original parametrization has typical member θ with density $p(\cdot; \theta)$, for any fixed one-to-one transformation $\phi = \psi(\theta)$ (which need not even be measurable, though in practice ψ will not only be measurable but even analytic) the same density may be represented with the parameter ϕ as

$$q(\cdot; \phi) = p(\cdot; \psi^{-1}(\phi)) . \qquad (1)$$

Definition

A method of estimation is invariant under reparametrizations if for any one-to-one transformation $\psi(\theta) = \phi$ of the parameter, the method yields estimators $\hat{\theta}$ for θ and $\hat{\phi}$ for ϕ such that $\hat{\phi} = \psi(\hat{\theta})$.

From (1) we have the following

Result

The method of maximum likelihood is invariant under reparametrizations.

It should be noted that invariance to reparametrizations is shared by a large class of methods for the discrete case, such as the method of minimum chi-squared, the method of minimum modified chi-squared, etc. In fact, invariance under reparametrizations is readily seen to hold in the identically and independently distributed case for all methods which are based on a criterion function (or an estimating equation) when the latter depends on θ through the set of likelihood elements $\{ q(x_i;\theta) . i = 1, 2, \ldots, n \}$. On the other hand several other methods, foremost among which is the method of moments, do not enjoy the property of invariance.

The value of the property is debatable: on the one hand one would want to keep open the possibility of using a different parametrization with a given situation, on the other hand the scope of allowable transformations is much too wide, since we are quite unlikely to use parametrizations which are not some smooth transformation of the original parametrization. Also, where an element of symmetry between the parameter space Θ and the space of observations Ξ is present in the problem (as discussed in the next subsection under cogredience), it seems that one would want to limit even more stringently the possible reparametrizations

to those which do not destroy this symmetry.

Note on Many-to-one Transformations of the Parameter

A more general result than the one just stated has been claimed by Zehna (1966). The method of maximum-likelihood is claimed to be invariant under any transformation ψ of the parameter, even when ψ maps different values of θ into the same value. Two objections can be made to this. First, as noted by Berk (1967), the likelihood function of which $\psi(\hat{\theta})$ is purported to be the maximum, does not in general correspond to the likelihood function of any random variable. Second, because of this fact the proof of the statement must contain a definition and so an 'induced' likelihood function $M(\phi) = \sup \{ p(x; \theta) : \phi = \psi(\theta) \}$ is introduced; on this definition hangs the result.

The justification given by Berk for calling $\psi(\hat{\theta})$ the MLE is that the use of the transformation ψ singles out a subset of the parameter space Θ in much the same way that considering the i-th component of the parameter θ focusses attention on a subset of Θ. If one feels comfortable in saying that the MLE for the mean μ of the two-parameter Gaussian is $\hat{\mu}$, when the full parameter is $\theta = (\mu, \sigma^2)$, then there is not too much harm in calling $\psi(\hat{\theta})$ the MLE of $\psi(\theta)$. With the Bernoulli distribution with standard parameter π, however, the variance $\psi(\pi) = \pi(1 - \pi)$ has the same dimensionality as the parameter, π, and in order to apply the above justification it appears one must consider a reparametrization such as

$$\phi(\pi) = (\pi(1 - \pi), (\pi < \tfrac{1}{2}))$$

so that the transformed parameter space is no longer a continuous subset of Euclidean space.

3.2.3 COGREDIENCE

Many parent parametric families for univariate x are of the type $p(x;\ (\mu, \sigma)) = p_o\left(\frac{x - \mu}{\sigma}\right)$, where we speak of μ as being a location parameter and of σ as a scale parameter. When a model assumes identically and independently distributed observations from such a family, there exists an intimate relationship between the data and the parameter $\theta = (\mu, \sigma)$. The latter is no longer a mere index for the family of distributions, it also has a 'physical' meaning and it seems reasonable to expect that the method of estimation being used will respect this relationship. The notion of location-and-scale model can be generalized to that of cogredience.

Definition

A cogredience model $\{ x;\ p(x;\ \theta);\ \theta \in \Theta \}$ is one where there exists a group G of transformations on Ξ and a group $\bar{G}$ of transformations on Θ such that

for all $g \in G$, there exists a $\bar{g} \in \bar{G}$ so that

$y = g(x)$ has density $p(\ g(x);\ \ \bar{g}(\theta)\ )$;

$\bar{g}$ is said to be the transformation induced by g. An estimator $\hat{\theta}$ is said to be cogredient under this model if $\bar{g}(\hat{\theta}) = \hat{\phi}$, where $\hat{\theta}$ is the estimator of θ based on the original data x and $\hat{\phi}$ is the estimator of $\phi = \bar{g}(\theta)$ based on the transformed data $g(x)$.

The transformations in G must have the property that for all $g \in G$ the measure $(\lambda g^{-1})(A) = \lambda(g^{-1}(A))$ be such that the dominating measure λ is absolutely continuous with respect to λg^{-1} (see Lehmann, 1959, p.252).

'Invariance' is the standard term for the notion we have termed 'cogredience'. The term 'cogredience' was used by Lehmann in a set of lecture notes (1950), although his book (1959) uses the standard term. We have preferred to avoid describing yet a third form of invariance in this section by the same same. The word 'cogredience' also has some appeal in the context of point estimation in that it serves as a reminder of the fact that changes are being made on two entities, and that a change on one of the two entities entails a corresponding change in the other.

Result

The maximum-likelihood estimator is cogredient under any cogredience model.

The proof of the result can proceed in one of two simple ways: either by combining the results of subsections 3.1.1 and 3.1.2, or directly, by noting that the likelihood function for the original and for the transformed parameters must be proportional, since the densities of corresponding data x and $g(x)$ are equal for corresponding parameters. The direct approach has this advantage: with it, it is clear that it refers to a property of the estimator itself, rather than to a property of the method of estimation.

The more general situation, where the transformed model is not the original one, but the transformation is sufficient, leads to a similar

invariance property for the method of maximum likelihood. However, not all data transformations which may reasonably be said to induce a transformation on the parameter could be considered will have the invariance property for the method of maximum likelihood.

An example of the more general kind of transformation has been considered by Dudewicz (1971). His discussion is in terms of a k-variate Gaussian distribution but we will restrict our attention to the bivariate situation. Let $x = (x_1, x_2)$ be distributed bivariate Gaussian with mean $\mu = (\mu_1, \mu_2)$ and identity covariance matrix. Let t be the transformation which orders the components in increasing order $t(x) = y = (y_1, y_2)$, $y_1 = \min(x_1, x_2)$, $y_2 = \max(x_1, x_2)$. Then y has density:

$$p(y; \mu) = [\, \varphi(y_1 - \mu_1)\, \varphi(y_2 - \mu_2) + \varphi(y_2 - \mu_1)\, \varphi(y_1 - \mu_2) \,] \; (\, y_1 < y_2 \,)$$

where $\varphi(\cdot)$ is the density of Gauss $(0,1)$.

Here, it seems natural to apply the same transformation to the parameter μ, since when the order of the data is lost, the ordering of the parameter components loses its meaning and the parameter is no longer identifiable, whereas the transformed parameter $t(\mu) = \nu = (\nu_1, \nu_2) = (\min(\mu_1, \mu_2), \max(\mu_1, \mu_2))$ is identifiable. Consider estimation of ν based on a single observation y. (A less appealing situation would have y be the ordered vector of averages of n observations; it is not clear that such a situation is realistic, however.) Dudewicz shows that the MLE for ν based on the ordered data pair y is:

$$\hat{\nu} = (\bar{x}, \bar{x}) \; (\, d \leq \sqrt{2} \,) + (\bar{x} - s, \bar{x} + s) \; (\, d > \sqrt{2} \,)$$

where $d = y_1 - y_2$, $\bar{x} = (x_1 + x_2)/2$, $s = \frac{\varepsilon_o}{2d}$ and ε_o is the unique positive root of $\varepsilon \coth(\varepsilon/2) = d^2$ when $d > \sqrt{2}$.

Thus except when $y_1 = y_2$, $\hat{\nu} \neq t(\hat{\mu})$ and the MLE, or rather the method of maximum likelihood, is not invariant under this type of double transformation.

3.2.4. ESTIMATION IN SEGMENTED MODELS

It seems appropriate to close this section, and this chapter, with a simple but interesting property which is related to both invariance and sufficiency.

Suppose that the data come naturally in two or more segments: $x = (y_1, y_2, \ldots, y_m)$, where the y_i may be multidimensional, not necessarily of the same dimension. The model $\{x; p(\cdot;\theta); \theta \in \Theta\}$ could be termed a segmented model when the y_i are jointly (stochastically) independent, and where the marginal distribution of y_i depends on a subparameter θ_i (not necessarily of the same dimension for all i) in such a way that the various θ_i are functionally independent, i.e., θ can be reparametrized to $(\theta_1, \theta_2, \ldots, \theta_m)$, which ranges over $\Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_m$. It is then possible to write

$$p(x;\theta) = \prod_{i=1}^{m} p_i(y_i;\theta_i) \quad (\theta_i \in \Theta_i)$$

for appropriate densities p_i.

Now, for all samples, the likelihood function factors into functionally independent segments, and the MLE for the subparameter θ_i will be a function of y_i only, and will not depend on any other segment

$$(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$$

of the data. *In other words, in a segmented model, the MLE will use only the data from the appropriate segment of the data to estimate the corresponding subparameter.*

A simple illustration of a segmented model is one where the data are taken to be n independent observations of a k-dimensional random variable with multivariate Gaussian distribution $\text{Gauss}(\mu, \sigma_0^2 I)$, where σ_0 is known and the paramater space is a Cartesian product of m subsets of the real line. (Typically, the parameter space is the full k-dimensional Euclidean space.) The MLE $\hat{\mu}$ is such that its i-th component only involves the i-th component of the observations. It thereby avoids the kind of objection that is sometimes leveled against Stein-type estimators, that they mix possibly incommensurate units. (See Efron & Morris, 1973, for a reference to a similar objection.)

A segmented model is an extreme case of a model which admits an 'S-sufficient' statistic (see Barndorff-Nielsen, 1978, p. 50). In this more general situation, there exist statistics t_1 and t_2 and a reparametrization $\theta \to (\theta_1, \theta_2)$, $\Theta \to \Theta_1 \times \Theta_2$ such that

$$p(x;\theta) = p_1(t_1;\theta_1)\, p_2(t_2;\theta_2|t_2) \qquad (2)$$

($p_2(t_2; \theta_2|t_2)$ being the density of a conditional distribution). Then t_1 is said to be 'S-sufficient' for θ_1 (and 'S-ancilliary' for θ_2). In (2), the MLE for the subparameter θ_1 would depend on x through t_1 only (although it would not in general be sufficient for θ_1).

CHAPTER 4 APPLICABILITY

This chapter surveys some of the factors which limit the applicability of maximum-likelihood estimation. We begin by studying some conceptual difficulties; this is followed by a survey of problems relating to the computational aspect of estimation; finally, the question of making distributional statements about the parameter is considered.

SECTION 4.1 CONCEPTUAL DIFFICULTIES

4.1.1 EXISTENCE

A fundamental difficulty is that, even when it is possible to specify a likelihood function, that function may fail to attain its maximum on the specified parameter space Θ.

There are situations where the parameter space is 'unnaturally' restricted, such as when it is specified that the standard parameter π of a Bernoulli distribution cannot equal $\frac{1}{2}$, or that the mean of a Gaussian distribution must be strictly positive. In such situations the likelihood function may take its maximum at one of the excluded values, so that the method fails to produce an estimator on the proper range.

A similar situation will occur when it is desired to exclude from Θ those values of θ which correspond to degenerate distributions. As an example, take a Bernoulli model with n observations and standard parameter π. When all n observations are equal to zero, the MLE is

$\hat{\pi} = 0$, which corresponds to a distribution concentrated at zero. For some purposes, such as where the estimate is to be used in a simulation study to produce samples similar to the one which was observed, such an estimate is not reasonable, and some other method of estimation must be used. (See Arnold, 1972, and Schafer, 1976 for some alternatives.)

Degenerate distributions are even more conceptually troublesome when the observations are continuous, since a degenerate distribution is not absolutely continuous with respect to Lebesgue measure. The likelihood function is typically unbounded in such cases. (Barnard, 1974, offers an interesting interpretation of such a situation.) It will often happen that the likelihood function is unbounded only for sets of data whose probability is zero under all models: for example, the Gaussian model with parameter (μ,σ^2) will only have an unbounded likelihood function (corresponding to a degenerate distribution) when all observations are equal; this possibility can be discounted by the fact that such an event has zero probability for all values of (μ,σ^2). However, there are models where the MLE corresponds to a degenerate distribution with probability greater than zero: an example is the lognormal model whose parent distribution has density:

$$q(x;(\mu,\zeta,\sigma)) = [(x-\mu)\sqrt{2\pi}\sigma]^{-1}\exp\{-\tfrac{1}{2}[\log(x-\mu)-\zeta]^2/\sigma^2\} \quad (\mu < x).$$

It can be shown (Hill, 1963) that for <u>any</u> sample, the likelihood is unbounded in the neighbourhood of the point $(\mu,\zeta,\sigma) = (x_{(1)}, -\infty, \infty)$ (where $x_{(1)}$ is the smallest observed value in the sample), the MLE corresponding therefore to the degenerate distribution concentrated at $(x_{(1)}, -\infty, \infty)$.

The last example may also serve to highlight the fact that, for the parametrization being considered, the MLE may be a point at infinity: such points may, or may not, be difficult to interpret, depending on the circumstances. Also, while singularities in the likelihood function can be removed by taking account of the discrete structure of the data (see later, subsection 4.1.3), the same is not true for points at infinity: the Bernoulli model with the exponential-family 'natural' parametrization $\theta = \log[\pi(1-\pi)^{-1}]$ will yield an estimate $\hat{\theta} = -\infty$ when all observations are zero.

4.1.2 ROBUSTNESS

A robust estimator is one which will still be roughly on target when the true model is somewhat different from the assumed model. It is rather difficult to make general statements about the robustness of the MLE, since each particular model would have to be considered separately against an appropriate set of alternatives. In general, though, the MLE ought not to be supposed to be robust, since its derivation makes such explicit use of the assumed model. Indeed, in the most common studies (e.g., Andrews, Bickel, Hampel, Huber, Rogers & Tukey, 1972) the MLE serves as a baseline, as a supposedly non-robust estimator against which other estimators are compared.

The notion of robustness seems to require a definition of parameter which is rather different from the one we have adopted here. Although Huber (1972) mentions the possibility of using a parametric family as a class of alternative distributions, a more fruitful approach

is to use more general classes of alternatives which are not all of the same parametric family. In such situations one should really speak of the subparameter of interest which might be, say, the median of the distribution (the other subparameters being 'incidental'). Furthermore, it is difficult to speak in this context of the parameter as indexing a family of distributions; at most, it indexes that aspect of the distribution which is felt to be relevant, such as its 'centrality' in the case of robust estimators of location.

4.1.3 DISCRETE STRUCTURE

Our last conceptual difficulty is that the observations can never be assumed to be continuous. Measurements can only be carried out to a definite number of decimal places, so that the variable which is effectively observed is discrete, and it may be argued that the model should properly be modified to account for the discrete structure of the data.

Thus, when the model is that observations are those of a Gaussian variate, one should really specify that the observations will be grouped so that $x_1 = 0.0$ will be reported whenever the true, underlying variate y_i is, e.g., in the range -0.05 to +0.05. The likelihood function for a set of observations $x_1, \ldots, x_n$ is then:

$$p(x;\ \mu,\sigma) = \prod_{j=-\infty}^{\infty} \left[\Phi\left(\frac{0.1j + 0.05 - \mu}{\sigma} \right) - \Phi\left(\frac{0.1j - 0.05 - \mu}{\sigma} \right) \right]^{n_j}, \quad (1)$$

where n_j is the number of y's in the range $0.1j - 0.05$ to $0.1j + 0.05$ (equivalently, the number of x's equal to $0.1j$) and $\Phi(\cdot)$ is the cumulative distribution function of the standard Gauss (0,1).

In a sense, the usual likelihood function.

$$p(x;\ \mu,\sigma) = \prod_{i=1}^{n} \varphi\left(\frac{x_i - \mu}{\sigma}\right)$$

(with φ denoting the density of Gauss $(0,1)$) can only be regarded as an approximation to the true likelihood function (1), and the standard MLE $(\hat{\mu},\hat{\sigma}^2) = \left(\bar{x},\ \frac{\Sigma(x_i - \bar{x})^2}{n}\right)$ only approximately maximizes (1).

Very often, the difference between the approximate MLE and the true one is quite negligible. In some models, however, the contrast is striking: for example, the likelihood function for the three-parameter lognormal model usually attains its maximum at an interior point of the parameter space, when account is taken of the discrete structure (see discussion in Barnard, 1966, and Kempthorne, 1966).

An extreme view of discrete structure would force us to always compute the MLE on the basis of a discretized model. However, a method of estimation should be computationally tractable as well as being conceptually good, and the extreme view would seem to exclude computationally simple MLE's in most models with an underlying, 'unobserved', continuous variate.

SECTION 4.2 COMPUTATIONAL DIFFICULTIES

4.2.1 AVAILABILITY OF A DENSITY

In most cases encountered, the likelihood function is available in terms of elementary functions and its general behaviour can be deduced readily. We wish to point out here that such is not always so.

The Pareto-Lévy 'stable' laws form a rather attractive parametric family. In particular, they constitute the class of limiting distributions of quantities such as $\left(\sum_{j=1}^{n} x_j - b_n \right) / a_n$, for suitable sequences $\{a_n\}$ and $\{b_n\}$, when the random variables $\{x_j\}$ are independently and identically distributed. Stable distributions can therefore be used to model variables which are considered to arise from some sort of averaging process: this argument is but a slight generalization of one of the common arguments advanced in favour of Gaussian models, and it leads to a far richer class of distributions.

In its most general form, a stable distribution is indexed by a four-dimensional parameter $(\alpha, \beta, \gamma, \delta)$, such that the characteristic function of the distribution is:

$$f(u) = E \exp(iux) = \exp\{ i\delta\omega - \gamma|\omega|^{\alpha} \{ 1 + i\beta\omega \operatorname{sign} u \} \},$$

where $\omega = \omega(u,\alpha) = \tan(\pi\alpha/2)$ $(\alpha \neq 1)$ $+ \frac{2}{\pi} \log|u|$ $(\alpha = 1)$.

δ and γ are, respectively, location and scale subparameters, while β is related to the skewness of the distribution and α, the 'characteristic exponent', determines the moments which exist finitely: for $\alpha < 2$, all moments of order $\alpha_o < \alpha$ exist, and no moments of order $\alpha_o > \alpha$ may exist.

It would appear (Feller, 1971, p.581; Paulson, Holcomb & Leitch, 1975) that densities of stable laws in terms of elementary functions are known in the following situations only (γ and δ are arbitrary)

$\alpha = 2$, $\beta = 0$ Gaussian distribution

$\alpha = 1$, $\beta = 0$ Cauchy distribution

$\alpha = \frac{1}{2}$, $\beta = 1$ this distribution does not seem to have a standard name, but its density is known to be:

$$p(x, \delta) = \frac{1}{\sqrt{2\pi}} (x - \delta)^{-3/2} \exp\left(\frac{-1}{2(x-\delta)}\right), \quad x > \delta.$$

For general members of the class of stable distributions, therefore, the density for each separate value of x and of $(\alpha, \beta, \gamma, \delta)$ would have to be computed numerically, either by integrating the characteristic function or by using the infinite series representation for the density (Feller, 1971, p.583).

Neither of the above alternatives is very palatable and those references we have seen eschew maximum-likelihood estimation in favour of other methods (e.g., Fama & Roll, 1968; Press, 1972; Paulson, Holcomb & Leitch, 1975). The problem may not be computationally prohibitive (after all, most elementary functions are effectively approximated numerically by a truncated version of the appropriate infinite series); however, one is rather uneasy about computing the maximum of a function whose shape one knows so little about.

4.2.2 SOLUTION

Although in many common models (Gaussian, one-parameter exponential, binomial, etc.) the MLE can be determined explicitly as a known function

of the observations, it is generally the case that the MLE can only be determined implicitly, and that its value must be obtained by numerical procedures. This is particularly true when the underlying distribution is conceived to be continuous but where it is decided to take account of the discrete structure of the observations.

It would be outside the scope of our survey to discuss all the numerical procedures which can be used to obtain the MLE. We merely note that the advent of computers has not entirely removed computational considerations from the statistician's purview.

To show that even a rather 'nice' looking density can produce a 'bad' likelihood function, consider the model of identically and independently distributed continuous observations from a parent distribution with density:

$$p(x;\theta) = \frac{2}{a}\frac{x}{\theta}\,(0 \le x \le \theta) + \frac{2}{a}\frac{a-x}{a-\theta}\,(\theta < x \le a) \qquad (1)$$

where a is a known constant. (1) is merely a triangular-shaped density with support on $(0,a)$ and mode at θ. Oliver (1972) shows that the likelihood function corresponding to (1) is continuous and piecewise convex. The convex 'pieces' are joined at the values $\theta = x_1, \ldots, x_n$ to form cusps. Thus any solution to the likelihood equation must lead to a local minimum, and the global maximum must be sought among the values $\theta = x_1, \ldots, x_n$. (Mantel, 1972, mentions an even more pathological example where the 'cusps' have infinite height.) It would appear to be a general rule that the computational aspect of finding the MLE must be carefully scrutinized, whenever the density is defined piecewise and whenever its carrier depends on the value of the parameter.

4.2.3. UNIQUENESS

Whenever the MLE cannot be obtained explicitly, it is of interest to determine whether the likelihood function has several modes. When a unimodal likelihood function cannot be assumed, special care must be exercised to ensure that the computational procedure selects the correct mode. A classic example of a 'smooth' multimodal likelihood function is provided by the model which assumes independently and identically distributed observations of a Cauchy variate with known scale (=1), and unknown median θ. The likelihood function is

$$p(x;\ \theta) = \frac{1}{\pi^n} \prod_{i=1}^{n} [\ 1 + (x_i - \theta)^2]^{-1} \qquad (2)$$

a rational function with no singularities on the real axis. Barnett (1966) showed in a simulation study that it is fairly common for (2) to be multimodal.

Another rather extreme case of multimodality is provided by the model with parent distribution uniform on $(\theta, \theta+1)$; the likelihood function is

$$p(x;\ \theta) = (\ x_{(1)} \geq \theta\ )(\ x_{(n)} \leq \theta + 1\ ) \qquad (3)$$

(where $x_{(1)}$ and $x_{(n)}$ are respectively the smallest and largest observations) and the maximum is attained at all

$$\theta \in [\ x_{(n)} - 1, x_{(1)}\ ].$$

Several approaches for proving the unimodality of the likelihood function are possible. The simplest approach is to show that the logarithm of the likelihood function is strictly concave. (Actually, it is sufficient

to show that some strictly monotone function of the likelihood function is strictly concave; the logarithm is usually the most convenient transformation to use, however.) It can be shown, for example, that in exponential families with the natural parametrization, the logarithm of the likelihood function is a strictly concave function, provided that the carrier of the densities is not concentrated on a proper affine subspace of k-dimensional Euclidean space, where $k = \dim \Theta$ (Barndorff-Nielsen, 1978, pp.103, 140). Therefore in exponential families the likelihood function is unimodal.

We note that unimodality does not in general hold true of curved exponential families: multimodality occurs in the bivariate Gaussian when only the correlation coefficient is to be estimated (Kendall & Stuart, 1973, p.40) or when the correlation coefficient is zero, the variances unknown and possibly distinct and both variates have unknown common mean (Fields, Kramer & Clunies-Ross, 1962).

When it is not possible to transform the likelihood function into a concave function, it may still be helpful to show that all likelihood sets

$$C(\theta_o) = \{ \theta : p(x; \theta) \geq p(x; \theta_o) \}$$

are convex (the likelihood function is then said to be pseudo-concave). Pseudo-concavity will not of itself guarantee unimodality, since the absolute maximum could still be attained on a connected set as it is in (3) above. However, it is sometimes possible to rule out the possibility of the likelihood function 'flattening out' to a plateau (see Antle, Klimko & Harkness, 1970, for one instance).

When concavity or pseudo-concavity cannot be proved, it is sometimes possible to determine the unimodality of the likelihood function by showing that only local maxima are possible. Two cases must be distinguished here. When the parameter is one-dimensional and the second derivative of the likelihood function exists everywhere, one can use Rolle's theorem to show that when the likelihood has two local maxima, there must be a local minimum between them. Therefore, if it is possible to show that:

$$\frac{d^2 p(x;\theta)}{d\theta^2} < 0$$

— for all θ such that $\frac{dp(x;\theta)}{d\theta} = 0$,

we will have shown unimodality of the likelihood function.

However, when the parameter has dimension greater than one, the situation is quite different. It has been noted by Tarone & Gruenhage (1975) that there exist smooth functions of two or more variables, such that there exist local maxima and no other turning points. It is true that the function exhibited by Tarone & Gruenhage is not known to be a likelihood function, but we know of no argument which would guarantee that a likelihood function cannot 'misbehave' in this way. It is rather unfortunate that on more than one occasion, the multidimensional case has been unjustifiably treated like the one-dimensional (an instance is in Huzurbazar, 1949, also in Kendall & Stuart, 1973, p.56).

Mäkeläinen, Schmidt & Styan (1976) have developed a rigorous criterion for the multidimensional case. A sequence of points $\{\theta_m\}$ in the parameter space Θ is said to converge to the boundary when: (1) it converges properly to a point on the finite boundary of Θ, or (2) $\|\theta_m\| \to \infty$.

The likelihood function $p(x; \cdot)$ is said to be 'constant on the boundary' whenever:

$$\lim_{m \to \infty} p(x; \theta_m) = c$$

for some constant c and for all sequences which converge to the boundary. Assume that:

(1) the likelihood function is constant on the boundary,

(2) the parameter space is a connected open subset of k-dimensional Euclidean space and

(3) the logarithm of the likelihood function is twice differentiable on Θ with negative-definite matrix of second derivatives $\nabla\nabla^T \log p(x; \theta)$ at all points θ such that $\nabla \log p(x; \theta) = 0$.

Under conditions (1), (2) and (3), Mäkeläinen, Schmidt & Styan (1976) show that the likelihood function must have a unique maximum. Thus it is sufficient to show constancy on the boundary in addition to nonexistence of saddle points and local minima, in order to show unimodality.*

Finally, a special method involving the partial solution of the likelihood equations has been mentioned by Cox (1976). In his method, at the k-th step the original parameters $\theta_1, \ldots, \theta_{k-1}$ have already been expressed in terms of $\theta_k, \ldots, \theta_q$. The k-th pivot is defined as $\frac{\partial^2 \log p_k(x; \theta_k, \ldots, \theta_q)}{\partial \theta_k^2}$. The likelihood equation $\frac{\partial \log p_k(x; \theta_k, \ldots, \theta_q)}{\partial \theta_k} = 0$ is then solved for θ_k

* The use of the criterion is illustrated in Appendix B, where it is used to repair an existing proof of unimodality.

in terms of $\theta_{k+1}, \ldots, \theta_q$, and $\theta_1, \ldots, \theta_{k-1}$ are similarly reexpressed in terms of $\theta_{k+1}, \ldots, \theta_q$. If all the q pivots can be shown to be negative, the likelihood function must be unimodal. An earlier application of this method (without a general statement, though) can be found in Pike (1966) to prove the unimodality of the likelihood function for a sample from a two-parameter Weibull parent population.

SECTION 4.3 DISTRIBUTIONAL INFERENCE

4.3.1 ASYMPTOTIC APPROXIMATIONS

Point estimation may be regarded as one type of statistical inference, since it allows us to say something about the possible nature of the true model. However, an inference of the type which allows us to make confidence statements, significance tests and the like is more satisfying. In this section we will survey those aspects of what might be called distributional inference, which are related to MLEs or to the likelihood function. Our review will be too brief to do justice to the importance of the topic, as we will merely sketch a few possible approaches. The purely frequentist results which will be discussed in the first two subsections are all derived from the fact that, under regularity conditions, the MLE from a model with identically and independently distributed observations has a distribution which is asymptotically Gauss $(\theta, I(\theta)^{-1}/n)$ where $I(\theta)$ is the Fisher information matrix

$$I(\theta) = E_\theta[\, \nabla_\theta \log p(x;\theta) \, (\nabla_\theta \log p(x;\theta))^T \,].$$

The regularity conditions usually required are rather stringent, including differentiability of $\log p(x;\cdot)$ to the third derivative (Cramér, 1946, pp.500-501). LeCam (1970) relaxes these conditions to differentiability of $\log p(x;\cdot)$ in mean square, but his proof seems to be valid only for a one-dimensional parameter.

A simple example of a parent distribution which does not lead to an asymptotic Gaussian distribution is provided by the uniform distribution on $(0,\theta)$. It is rather easy to establish that the MLE $\hat{\theta} = \max(x_1,\ldots,x_n)$

is asymptotically exponentially distributed with mean θ/n.

Improved Approximations

One way of improving the accuracy of the Gaussian approximation (where it is available) has been developed by Haldane & Smith (1956). In the case of discrete distributions with bounded parameter space, they were able to provide approximate expressions for the third and fourth cumulants of the distribution of the MLE, in terms of the true parameter. Haldane & Smith were primarily interested in approximating the moments themselves, but it has been suggested by Kendall & Stuart (1973, p.50) that these cumulants could be used to obtain approximate confidence intervals for the parameter by way of the Pearson system of distributions.

A more fruitful way of manipulating the standard Gaussian approximation is to use quantities which are at least asymptotically pivotal, i.e., whose asymptotic distribution is the same for all values of the true parameter. One such asymptotic pivotal is

$$\frac{1}{n}(\hat{\theta} - \theta)^T I(\theta) (\hat{\theta} - \theta) ,$$

which, under the regularity conditions, has asymptotically a χ^2_k distribution (where $k = dim\theta$). Sprott (1975) considers several asymptotic pivotals in the one-dimensional case.

One might seek to improve such asymptotic pivotals by deriving other pivotals whose moments coincide even more closely with those of a Gaussian distribution. Bartlett (1953a, b) refines the score pivotal

$$\frac{d \log p(x; \theta)}{d\theta} [I(\theta)]^{-\frac{1}{2}}$$

obtaining from it a quantity whose skewness vanishes asymptotically. Other

work along these lines has been done by Welch & Peers (1963) when θ is a location parameter.

'Normality' of the Likelihood Function

One drawback to the usefulness of any asymptotic result is the absence of any useful bound on the error involved, say, in the difference between the actual significance level attained by a statistic, and the level predicted by the asymptotic theory used to derive that statistic. Otherwise stated, when is the sample size 'large' enough? One appealing approach uses the notion of 'normality' of the likelihood function. Briefly, if the asymptotic distribution of $\hat{\theta}$ were in fact Gaussian, the observed likelihood function would of course be a bell-shaped function exactly proportional to the density of a Gaussian variate. Therefore it may be hoped that when the observed likelihood function is approximately bell-shaped, the asymptotic approximation to the distribution of $\hat{\theta}$ is fairly accurate (Sprott & Kalbfleisch, 1969). Of course it has been pointed out by the proponents of the argument that it is not legitimate to infer from an observed 'normal' likelihood function that the distribution is in fact Gaussian.

Perhaps the most satisfying justification for the 'normality' criterion is a weak version of the likelihood principle. The strong likelihood principle, as promoted by Birnbaum (1962), would (loosely speaking) have statistical inferences based on the observed likelihood function, without any reference to those elements of the design of the experiment which are not reflected in the likelihood function. The weak version of the principle is similar but would exclude situations where, for example, a cogre-

dience structure exists (Barnard & Sprott, 1971).

We note, however, that even the weak likelihood principle departs from strict frequentist principles. Thus the distribution of the MLE $\hat{\pi}$ of the standard Bernoulli parameter is a reflection of all possible samples, those which are nicely spread out (yielding a close approximation to a Gaussian density) and those that are more extreme, such as happens when all observations are equal to zero or to one (the likelihood function for such a sample is not at all 'normal').

If one accepts the notion of 'normality' for likelihood functions, it is possible to derive techniques which appear to be useful in refining the asymptotic approximation. Thus one may seek a parametrization which would make the MLE more nearly Gaussian. Sprott (1973) has developed numerical criteria for judging whether a given reparametrization would improve things. The idea of 'transforming to normality' was also considered by Anscombe (1964) (but with a Bayesian justification). See also Mitchell (1962).

Linearity and Adequacy of Fit

A further drawback of asymptotic approximations is that these, of necessity, tend to fit the 'centre' of the distribution rather closely, but that the approximation may be poor at the tails (where, of course, interest is likely to centre).

In a slightly similar vein, it has been suggested (Sprott, 1975) that in deciding whether to refine an approximation, account be taken of whether the resulting approximation would be more, or less, linear in the parameter. A non-linear approximation may lead to unsatisfactory results in the tail areas, such as a confidence interval which is not properly

contained in the parameter space.

4.3.2 LIKELIHOOD SETS

We now consider inference based on likelihood sets. It will be assumed that $p(x; \cdot) < \infty$ and the sets will for convenience be denoted by:

$$C(r) = C(r; x) = \{ \theta : p(x; \theta) \geq r p(x; \hat{\theta}) \}, \qquad (1)$$

where $r \in (0,1)$. It is appealing to use likelihood sets in distributional inference, first because in the most familiar models where uniformly most powerful tests are available, these are based on the likelihood ratio, so that one might hope that the 'good' properties will hold, in attenuated form, for more general situations. A second reason for using likelihood sets is that these are the natural extension, so to speak, of the MLE; if one accepts Birnbaum's characterization (1964) of a point estimate as a confidence interval with 0% confidence coefficient, then the MLE is the 0% confidence interval belonging to the family of confidence intervals which are likelihood sets.

Frequentist Approach

When the conditions for the MLE to have an asymptotic Gaussian distribution are satisfied, the likelihood ratio is asymptotically pivotal and likelihood sets may readily be used as approximate confidence intervals for the parameter, with approximate content determined by r. Thus, setting $r = e^{-2}$ one obtains that $C(r)$ has roughly 95% content (the asymptotic level being $P(\theta \in C(r) \mid \theta) = 0.954$).

Hudson (1971) has reviewed the performance of likelihood sets for

a number of models. Aside from assessing the true situation in specific models, his review is a useful reminder that in general, the actual confidence content of a likelihood set will depend, not only on r, but also on the sample size and the true parameter value.

A result for ensuring that the content is independent of θ has been given by Spjøtvoll (1972). The result is available only where a cogredience structure is present, and where mild regularity conditions are satisfied on either the group of transformations or the likelihood function. It is even shown that not only do the $C(r)$ have content independent of θ, but also that considered as confidence sets, they are unbiased. A stronger result by Joshi (1970) has that for a one-dimensional location parameter and under regularity conditions, the $C(r)$ are minimax among all confidence sets with the same confidence coefficient, i.e., they have smallest expected (Lebesgue) volume.

Likelihood Inference

We close this subsection by mentioning a mode of reasoning based on the likelihood function which departs frankly from standard frequentist theory. The basic idea is that r in (1) can be used as a measure of plausibility for a value of θ on the boundary of $C(r)$. It would be too long to detail the history of this, but it may be noted that something along those lines seems to have been advocated early by Fisher (in papers such as 1921 and 1925, with a more explicit use in 1959, p.74). (Edwards (1972) has a more complete theory of the subject.) We may also note that the notion has some support from the Bayesian viewpoint, since when an (improper) prior distribution is assumed for θ, r is just the

posterior probability of $C(r)$ (Anscombe, 1961). We retain from critics of the approach (Cox, 1958; Plackett, 1966) that when several similar models are considered, to a given value of r will correspond different confidence coefficients, according in part to the dimension of the parameter.

4.3.3 STOCHASTIC ORDERING

We may end this discussion on a positive note by mentioning a property of the probability distribution of the MLE in certain models. The result is due to Pfanzagl (1976).

Result

If the MLE, $\hat{\theta}$ of a one-dimensional parameter θ exists uniquely for almost all samples and is measurable, and if the family of distributions $\{ p(\cdot; \theta); \theta \in \Theta \}$ for a univariate observation has constant carrier, non-decreasing likelihood ratio ($\theta^* > \theta \Rightarrow p(x; \theta^*)/p(x; \theta)$ is a non-decreasing function of x) and is stochastically ordered ($\theta^* > \theta \Rightarrow \int_{-\infty}^{x} p(y; \theta^*) d\lambda(y) \geq \int_{-\infty}^{x} p(y; \theta) d\lambda(y)$), then the family of distributions of the MLE is stochastically ordered also.

Note: the result is valid for observations on any ordered measurable space, and an analogous result is indicated in the case of independently but not identically distributed observations.

When $\frac{\partial^2 \log p(x;\theta)}{\partial\theta\partial x}$ exists, it is a necessary and sufficient condition for $\{ p(\cdot;\theta);\ \theta \in \Theta \}$ to have monotone likelihood ratio that $\frac{\partial^2 \log p(x;\theta)}{\partial\theta\partial x} \geq 0$ for all θ and all x (Lehmann, 1959, p.111). The above result is therefore not available in many models. It may be noted that when

$$p(x;\theta) = \exp[\, t(x)\ \phi(\theta) - \kappa(\theta) - g(x)\ ]$$

(i.e., in an exponential family with *dim* ϕ = *dim* θ = 1) the condition reduces to the requirement that ϕ be a monotone function of θ.

CHAPTER 5 CONCLUSION

For the author, the principal attractiveness of maximum-likelihood estimation lies in the fact that its use can be attempted in most models, whether with discrete or continuous data, when the observations are independently and identically distributed, and when the model has some other, more complicated structure. It seems to come closer than any other method to being the 'portmanteau' method of point estimation. However, it cannot be used effectively in all models, as discussed in Chapter 4.

What, then, are the finite-sample properties of maximum-likelihood estimation? The method can properly be said to produce estimators, since MLEs are specific in the sense of Fisher consistency. Apart from Fisher consistency, MLEs have another reason for being called specific in models where a cogredience (or invariance) structure exists.

The invariance of MLEs to reparametrization is a rather significant property in any model where the parametrization is arbitrary.

On the other hand, the only kind of optimality property enjoyed by MLEs under closeness criteria would appear to be sensitivity, and the appeal of this notion is not immediate.

Even in models where it is possible to find a sufficient statistic of the same dimension as the parameter, the MLE need not be sufficient, though sufficiency is attained by the MLE in many examples of practical importance.

There is no universal, practicable way of making frequentist

probability statements about the parameter on the basis of the MLL. However, the approximate distributional theory is rather well-developed.

In general, it could be said that the maximum-likelihood estimation exploits the highly specific, parametric, non-robust aspects of the model. Its use will be most satisfactory where one is rather confident that the model adequately describes the situation which effectively gave rise to the data set under consideration.

REFERENCES

The list of references which follows consists largely of those workswhich are cited in the text. A comprehensive bibliography of maximum-likelihood estimation, covering articles up to 1970, appears in Norden (1972).

Andrews, D.F.; Bickel, P.J.; Hampel, F.R.; Huber, P.J.; Rogers, W.H.; & Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton Univ. Press, Princeton, N.J.

Anscombe, F.J. (1961). Estimating a mixed-exponential response law. *J. Amer. Statist. Assoc.*, 56, 493-502.

Anscombe, F.J. (1964). Normal likelihood functions. *Ann. Inst. Statist. Math. Tokyo*, 16, 1-19.

Antle, C; Klimko, L; & Harkness, W. (1970). Confidence intervals for the parameters of the logistic distribution. *Biometrika*, 57, 397-402.

Arnold, B.C. (1972). Some examples of minimum variance unbiased estimates. *Amer. Statist.*, 26, 34-36.

Bahadur, R.R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.*, 25, 423-462.

Barankin, E.W. & Maitra, A.P. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. *Sankhyā Ser. A.*, 25, 217-244.

Barnard, G.A. (1962). [Discussion of Rao (1962b)]. *J. Roy. Statist. Soc. Ser. B*, 24, 67-68.

Barnard, G.A. (1963). Some logical aspects of the fiducial argument. *J. Roy. Statist. Soc. Ser. B*, 25, 111-114.

Barnard, G.A. (1966). The use of the likelihood function in statistical practice. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, 1, 27-40, Univ. of California Press, Berkeley.

Barnard, G.A. (1973). Maximum likelihood and nuisance parameters. *Sankhyā Ser. A*, 35, 133-138.

Barnard, G.A. (1974). [Letter to the Editor]. *Amer. Statist.*, 28, 162.

Barnard, G.A. (1977). On ridge regression, and the general principles of estimation. *Utilitas Math.*, 11, 299-311.

Barnard, G.A. & Sprott, D.A. (1971). A note on Basu's examples of anomalous ancillary statistics. Pp.163-176 in: Godambe & Sprott (1971).

Barndorff-Nielsen, O. (1970). *Exponential Families: Exact Theory*. Aarhus Universitet Matematisk Institut, Various Publ. Series No.19, Aarhus, Denmark. [Substantially expanded in Barndorff-Nielsen (1978)].

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, England.

Barnett, V.D. (1966). Evaluation of the maximum likelihood estimator when the likelihood equation has multiple roots. *Biometrika*, 53, 151-165.

Barnett, V.D. (1973). *Comparative Statistical Inference*. Wiley, London.

Bartlett, M.S. (1953a). Approximate confidence intervals. *Biometrika*, 40, 12-19.

Bartlett, M.S. (1953b). Approximate confidence intervals II. More than one unknown parameter. *Biometrika*, 40, 306-317.

Barton, D.E. (1956). A class of distributions for which the maximum-likelihood estimator is unbiased and of minimum variance for all sample sizes. *Biometrika*, 43, 200-202.

Berk, R.H. (1967). [Review of Zehna (1966)]. *Math. Reviews*, 33, 342-343. [#1922].

Bhapkar, V.P. (1972). On a measure of efficiency of an estimating equation. *Sankhyā Ser. A*, 34, 467-472.

Bhapkar, V.P. (1973). A remark on the invariance of some estimators. *Amer. Statist.*, 27, 231-232.

Birnbaum, A. (1961). A unified theory of estimation, I. *Ann. Math. Statist.*, 32, 112-135.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Soc.*, 57, 269-326.

Birnbaum, A. (1964). Median-unbiased estimators. *Bull. Math. Stat. Fukuoka*, 11 (1 & 2), 25-34.

Bishop, Y.M.M.; Fienberg, S.E.; & Holland P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

Burkholder, D.L. (1961). Sufficiency in the undominated case. *Ann. Math. Statist.*, 32, 1191-1200.

Canfield, R.V. (1970). A Bayesian approach to reliability estimation using a loss function. *IEEE Trans. Reliab.*, 19, 13-16.

Carnap, R. (1962). *Logical Foundations of Probability*. Second Edition. Univ. of Chicago Press, Chicago. [First Edition, 1950].

Chentsov, N.N. [Čencov] (1964). Geometry of a manifold of probability distributions. *Soviet Math. Dokl.*, 5, 1282-1286.

Chernoff, H. (1976). The interaction between large sample theory and optimal design of experiments. In: Owen, D.B., ed., *On the History of Statistics and Probability*, 205-223.

Copas, J.B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika*, 62, 701-704.

Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357-372.

Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Cox, N.R. (1976). A note on the determination of the nature of turning points of likelihoods. *Biometrika*, 63, 199-201.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, N.J.

Curry, R.E. (1976). Sufficient conditions for the uniqueness of parameter estimates from binary-response data. *J. Mathematical Psychology*, 14, 72-90.

Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *C.R. Acad. Sci. Paris*, 200, 1265-1266.

Deutsch, R. (1965). *Estimation Theory*. Prentice-Hall, Englewood Cliffs, N.J.

Dudewicz, E.J. (1971). Maximum likelihood estimators for ranked means. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 19, 29-42.

Dutta, M. (1966). On maximum (information-theoretic) entropy estimation. *Sankhyā Ser. A*, 28, 319-328.

Edgeworth, F.V. (1909). On the probable error of frequency constants. (Addendum). *J. Roy. Statist. Soc.*, 72, 81-90.

Edwards, A.W.F. (1972). *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*. Cambridge Univ. Press, Cambridge, England.

Edwards, A.W.F. (1974). The history of likelihood. *Internat. Statist. Rev.*, 42, 9-45.

Efron, B. (1975). Biased versus unbiased estimation. *Advances in Math.*, 16, 259-277.

Efron, B. & Morris, C. (1973). Combining possibily related estimation problems. *J. Roy. Statist. Soc. Ser. B*, 35, 379-421.

Fama, E. & Roll, R. (1968). Some properties of symmetric stable distributions. *J. Amer. Statist. Assoc.*, 63, 817-836.

Farebrother, R.W. (1977). Modified estimators for the binomial parameter. [Letter to the Editor]. *Amer. Statist.*, 31, 98.

Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21, 826-838.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications. Volume II.* Second Edition. Wiley, New York.

Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York.

Fields, R.L.; Kramer, C.Y.; & Clunies-Ross, C.W. (1962). Joint estimation of the parameters of two normal populations. *J. Amer. Statist. Assoc.*, 57, 446-454.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80, 758-770. [Paper 2 in Fisher (1950)].

Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1, (4), 3-32.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, 222, 309-368. [Paper 10 in Fisher (1950)].

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.*, 22, 700-725. [Paper 11 in Fisher (1950)].

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. London Ser. A*, 144, 285-307. [Paper 24 in Fisher (1950)].

Fisher, R.A. (1935). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc.*, 98, 39-54. [Paper 26 in Fisher (1950)].

Fisher, R.A. (1950). *Contributions to Mathematical Statistics.* Wiley, New York.

Fisher, R.A. (1958). *Statistical Methods for Research Workers.* Thirteenth Edition. Hafner, New York. [First Edition, 1925].

Fisher, R.A. (1959). *Statistical Methods and Scientific Inference.* Second Edition. Oliver & Boyd, Edinburgh. [First Edition, 1956].

Flinger, M.A.; Policello, G.E.; & Singh, J. (1977). A comparison of two randomized response survey methods with consideration for the level of response protection. *Comm. Statist. - Theory Methods*, A6, 1511-1524.

Fraser, D.A.S. (1952). Sufficient statistics and selection depending on the parameter. *Ann. Math. Statist.*, 23, 417-425.

Fraser, D.A.S. (1963). On sufficiency and the exponential family. *J. Roy. Statist. Soc. Ser. B*, 25, 115-123.

Fraser, D.A.S. (1966). Sufficiency for regular models. *Sankhyā Ser. A*, 28, 137-144.

Ghosh, J.K. & Singh R. (1970). Estimation of the reciprocal of the scale parameter of a gamma density. *Ann. Inst. Statist. Math. Tokyo*, 22, 51-55.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31, 1208-1211.

Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63, 277-284.

Godambe, V.P. & Sprott, D.A. (eds.) (1971). *Foundations of Statistical Inference*. Holt, Rinehart & Winston - Canada, Toronto. [Proceedings of a Symposium held at the University of Waterloo, March - April, 1970].

Godambe, V.P. & Thompson, M.E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.*, 2, 568-571.

Gokhale, D.V. & Kullback, S. (1978). *The Information in Contingency Tables*. Dekker, New York.

Good, I.J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.*, 34, 911-934.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge Univ. Press, Cambridge, England.

Haldane, J.B.S. & Smith, S.M. (1956). The sampling distribution of a maximum likelihood estimate. *Biometrika*, 43, 96-103.

Hampel, F.R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 27, 87-104.

Han, C.-P. (1969). Maximum likelihood estimate in intraclass correlation model. *Technometrics*, 11, 833-834.

Harter, H.L. & Moore, A.H. (1966). Local maximum likelihood estimation of the parameters of three parameter lognormal populations with complete and censored samples. *J. Amer. Statist. Assoc.*, 61, 842-851.

Hartigan, J.A. (1967). The likelihood and invariance principles. *J. Roy. Statist. Soc. Ser. B*, 29, 533-539.

Higgins, J.J. (1977). Bayesian inference and the optimality of maximum likelihood estimation. *Internat. Statist. Rev.*, 45, 9-11.

Hill, B.M. (1963). The three-parameter lognormal distribution and Bayesian analysis of a source-point epidemic. *J. Amer. Statist. Assoc.*, 58, 72-84.

Hill, B.M. (1975). Aberrant behavior of the likelihood function in discrete cases. *J. Amer. Statist. Assoc.*, 75, 717-719.

Huber, P.J. (1972). Robust statistics: a review. *Ann. Math. Statist.*, 43, 1041-1067.

Hudson, D.J. (1971). Interval estimation from the likelihood function. *J. Roy. Statist. Soc. Ser. B*, 33, 256-262.

Huzurbazar, V.S. (1949). On a property of distributions admitting sufficient statistics. *Biometrika*, 36, 71-74.

Huzurbazar, V.S. (1976). *Sufficient Statistics: Selected Contributions*. Dekker, New York.

Jagers, P. (1977). A decent likelihood function is certainly minimal sufficient. *Scand. J. Statist.*, 4, 35-36.

James, W. & Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1, 361-379.

Jeffreys, H. (1960). An extension of the Pitman-Koopman theorem. *Proc. Cambridge Philos. Soc.*, 56, 393-395.

John, S. (1974a). Median-unbiased most acceptable estimates of Poisson, binomial and negative-binomial distributions. *Comm. Statist.*, 3, 1155-1159.

John, S. (1974b). Acceptability and statistical inference. *Biometrika*, 61, 285-290.

Johnson, N.L. & Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Houghton-Mifflin, New York.

Johnson, N.L. & Kotz, S. (1970a). *Distributions in Statistics: Continuous Univariate Distributions - 1*. Houghton-Mifflin, New York.

Johnson, N.L. & Kotz, S. (1970b). *Distributions in Statistics: Continuous Univariate Distributions - 2*. Houghton-Mifflin, New York.

Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

Joshi, V.M. (1970). Admissibility of invariant confidence procedures for estimating a location parameter. *Ann. Math. Statist.*, 41, 1508-1581.

Kalbfleisch, J.D. & Sprott, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. Ser. B*, 32, 175-208.

Kalbfleisch, J.G. & Sprott, D.A. (1974). Inferences about hit number in a virological model. *Biometrics*, 30, 199-208.

Kallianpur, G. & Rao, C.R. (1955). On Fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā*, 15, 331-342. [Corrigenda in *Sankhyā*, 16, 206].

Kempthorne, O. (1966). Some aspects of experimental inference. *J. Amer. Statist. Assoc.*, 61, 11-34.

Kempthorne, O. & Folks, L. (1971). *Probability, Statistics, and Data Analysis*. Iowa State Univ. Press, Ames.

Kendall, M.G. (1961). Studies in the history of probability and statistics, IX. Daniel Bernoulli on maximum likelihood. *Biometrika*, 48, 1-2.

Kendall, M.G. & Stuart, A. (1973). *The Advanced Theory of Statistics, Volume 2.* Third Edition. Griffin, London.

Konijn, H.S. (1963). A note on the non-existence of a maximum likelihood estimate. *Austral. J. Statist.*, 5, 143-146.

Koopman, B.O. (1936). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, 39, 399-409.

Kŏrezlioğlu, H. (1969). [Review of Kriz & Talacko (1968)]. *Math. Reviews*, 38, 1206-1207. [#6698].

Kriz, T.A. & Talacko, J.V. (1968). Equivalence of the maximum likelihood estimator to a minimum entropy estimator. *Trabajos Estadist.*, 19, (I/II), 55-65.

Kullback, S. (1968). *Information Theory and Statistics.* Dover, New York. [Original Edition, Wiley, 1959].

Lambert, J.A. (1972). Likelihoods with rounded data. *Austral. J. Statist.*, 14, 204-210.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, 41, 802-828.

Lehmann, E.L. (1950). *Notes on the Theory of Estimation.* Associated Students' Store, Univ. of California, Berkeley. [Notes recorded by Colin Blyth].

Lehmann, E.L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

Lindley, D.V. (1972). *Bayesian Statistics, a Review.* Society for Industrial and Applied Mathematics, Philadelphia. [Cover dated 1971].

Lindsey, J.K. (1974). Comparison of probability distributions. *J. Roy. Statist. Soc. B*, 36, 38-47.

Mäkeläinen, T; Schmidt, K.; & Styan, G.P.H. (1976). On the uniqueness of the maximum likelihood estimator of a vector-valued parameter. Discussion Paper No.140, Indian Statistical Institute, Delhi Campus. [Abstracted (1977) in *Canad. J. Statist.*, 5, 256].

Mandelbrot, B. (1962). The role of sufficiency and of estimation in thermodynamics. *Ann. Math. Statist.*, 33, 1021-1038.

Mantel, N. (1972). Another maximum likelihood oddity. [Letter to the Editor]. *Amer. Statist.*, 26, 45.

McCool, J.I. (1970). Inference on Weibull percentiles and shape parameter from maximum likelihood estimates. *IEEE Trans. Reliab.*, 19, 2-9.

Mitchell, A.F.S. (1962). Sufficient statistics and orthogonal parameters. *Proc. Cambridge Philos. Soc*, 58, 326-337.

Neyman, J. (1971). Foundations of behavioristic statistics. Pp.1-19 in: Godambe & Sprott (1971).

Norden, R.H. (1972). A survey of maximum likelihood estimation. *Internat. Statist. Rev.*, 40, 329-354.

Norden, R.H. (1973). A survey of maximum likelihood estimation, Part 2. *Internat. Statist. Rev.*, 41, 39-58.

Oliver, E.H. (1972). A maximum likelihood oddity. *Amer. Statist.*, 26, 43-44.

Paulson, A.S.; Holcomb, E.W.; & Leitch, R.A. (1975). The estimation of the parameters of the stable laws. *Biometrika*, 62, 163-170.

Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, 22, 142-161.

Pitman, E.J.G. (1936). Sufficient statistics and intrinsic accuracy. *Proc. Cambridge Philos. Soc.*, 32, 567-579.

Pitman, E.J.G. (1939). The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika*, 30, 391-421.

Plackett, R.L. (1966). Current trends in statistical inference. *J. Roy. Statist. Soc. Ser. A*, 129, 249-267.

Plante, A. (1971). Counter-examples and likelihood. Pp.357-371 in: Godambe & Sprott (1971).

Plante, A. (1976). On maximum likelihood estimation of the parameter of an ordered family of distributions. *Utilitas Math.*, 9, 113-122.

Ponnapalli, R. (1976). Deficiencies of minimum discrepancy estimators. *Canad. J. Statist.*, 4, 33-50.

Pratt, J.W. (1976). F.Y. Edgeworth and R.A. Fisher on the efficiency of maximum likelihood estimation. *Ann. Statist.*, 4, 501-514.

Press, S.J. (1972). Estimation in univariate and multivariate stable distributions. *J. Amer. Statist. Assoc.*, 67, 842-846.

Proschan, F. & Sullo, P. (1976). Estimating the parameters of a multivariate exponential distribution. *J. Amer. Statist. Assoc.*, 71, 465-472.

Rao, C.R. (1962a). Apparent anomalies and irregularities in maximum likelihood estimation (with discussion). *Sankhyā Ser. A*, 24, 73-101. [Reprinted from *Bull. Inst. Internat. Statist.*, 38, 439-453; 38, 193-214; dated 1961].

Rao, C.R. (1962b). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser. B*, 24, 46-72.

Rényi, A. (1961). On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1, 547-561, Univ. of California Press, Berkeley.

Savage, L.J. (1972). *The Foundations of Statistics.* Second Edition. Dover, New York. [First Edition, Wiley, 1954].

Savage, L.J. (1976). On rereading R.A. Fisher. *Ann. Statist.*, 4, 441-500. [Posthumous paper edited by J.W. Pratt.]

Schafer, R.E. (1976). Modified estimators for the binomial parameter. *Amer. Statist.*, 30, 98-100.

Sethuraman, J. (1961). Conflicting criteria of 'goodness' of statistics. *Sankhyā Ser. A*, 23, 187-190.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27, 379-423 & 623-656.

Silverstone, H. (1957). Estimating the logistic curve. *J. Amer. Statist. Assoc.*, 52, 567-577.

Silvey, S.D. (1975). *Statistical Inference*. Chapman & Hall, London. [Original Edition, Penguin, 1970].

Simon, G. (1973). Additivity of information in exponential family probability laws. *J. Amer. Statist. Assoc.*, 68, 478-482.

Solari, M.E. (1969). The "maximum likelihood solution" of the problem of estimating a linear functional relationship. *J. Roy. Statist. Soc. Ser. B*, 31, 372-375.

Spjøtvoll, E. (1972). Unbiasedness of likelihood ratio confidence sets in cases without nuisance parameters. *J. Roy. Statist. Soc. Ser. B*, 34, 268-273.

Sprott, D.A. (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika*, 60, 457-465.

Sprott, D.A. (1975). Application of maximum likelihood methods to finite samples. *Sankhyā Ser. B*, 37, 259-270.

Sprott, D.A. (1978). Gauss's contributions to statistics. *Historia Math.*, 5, 183-203.

Sprott, D.A. & Kalbfleisch, J.D. (1969). Examples of likelihoods and comparison with point estimates and large sample approximations. *J. Amer. Statist. Assoc.*, 64, 468-484.

Stigler, S.M. (1973). Studies in the history of probability and statistics XXXII, Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60, 439-445.

Stigler, S.M. (1976). [Discussion of Savage, 1976]. *Ann. Statist.*, 4, 498-500.

Sverdrup, E. (1967). *Basic Concepts in Statistical Inference: Laws and Chance Variations, Volume II*. North Holland, Amsterdam.

Tarone, R.E. & Gruenhage, G. (1975). A note on the uniqueness of roots of the likelihood equation for vector-valued parameters. *J. Amer. Statist. Assoc.*, 70, 903-904.

Tiao, G.C. & Box, G.E.P. (1973). Some comments on "Bayes" estimators. *Amer. Statist.*, 27, 12-14.

Tukey, J.W. (1960). Conclusions vs decisions. *Technometrics*, 2, 423-433.

Tusnády, G. (1968). Remark on K. Sarkadi's paper entitled: "Estimation after selection". *Studia Sci. Math. Hungar.*, 3, 381-382.

Wasan, M.T. (1965). Theory of modal unbiased estimation. [Abstract]. *Ann. Math. Statist.*, 36, 1324.

Wasan, M.T. (1970). *Parametric Estimation*. McGraw-Hill, New York.

Wedderburn, R.W.M. (1976). On the existence and uniqueness of maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63, 27-32.

Welch, B.L. (1939). On the distribution of maximum likelihood estimates. *Biometrika*, 31, 187-190.

Welch, B.L. & Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B*, 25, 318-329.

Whitehouse, H.L.K. (1969). *Towards an Understanding of the Mechanism of Heredity*. Second Edition. Arnold, London.

Zacks, S. (1966). Sequential estimation of the mean of a log-normal distribution having a prescribed proportional closeness. *Ann. Math. Statist.*, 37, 1688-1696.

Zacks, S. (1967). On the non-existence of a fixed sample estimator of the mean of a log-normal distribution having a prescribed proportional closeness. *Ann. Math. Statist.*, 38, 949.

Zacks, S. (1971). *The Theory of Statistical Inference*. Wiley, New York.

Zacks, S. & Even, M. (1966). The efficiencies in small samples of the maximum likelihood and best unbiased estimators of reliability functions. *J. Amer. Statist. Assoc.*, 61, 1033-1051.

Zehna, P.W. (1966). Invariance of maximum likelihood estimators. *Ann. Math. Statist.*, 37, 744.

APPENDIX A

CALCULATION OF THE TABLE IN SUBSECTION 2.1.3

THE MODEL

The situation assumed by Wasan (1970, pp.162-171) is as follows. A sample of n observations $z_1, \ldots, z_n$ is taken from a Weibull parent distribution with known shape parameter K, unknown scale parameter ϕ and zero threshold. The observations might represent the observed lifetimes of some type of electronic component under test. It is desired to estimate from the sample the reliability of a typical component, that is, the probability that it will fail before the specified mission time. Now a Weibull variate z has the same distribution as y^K, where y is exponentially distributed with mean $\theta = \phi^K$. We will therefore work with the simpler exponential distribution, with density

$$p(y;\, \theta) = \theta^{-1} \exp(-y/\theta) \quad (\; y > 0 \;).$$

Let the specified mission time be denoted y_o (representing y_o^K in the original Weibull scale). The reliability of one component is then

$$\rho = P_\theta(y > y_o) = \exp(-y_o/\theta). \tag{1}$$

When a system consists of m identical components arranged in series, so that the system fails as soon as one component fails, the reliability of the system as a whole is ρ^m.

ESTIMATION

The reliability ρ is a one-to-one function of θ, so that the MLE of ρ^m is simply $\hat{\rho}^m = \exp(-y_o m/\hat{\theta})$. Now the MLE in an exponential model is $\hat{\theta} = \Sigma y_i/n$ so it is easily seen that the pivotal quantity $x = \frac{\hat{\theta} n}{\theta}$ has a gamma distribution with scale parameter 1 and shape parameter n. Thus we have:

$$E\hat{\rho}^m = \int_0^\infty \exp\left(-\frac{y_o m n}{x}\right) \frac{x^{n-1} e^{-x}}{(n-1)!}\, dx \;;$$

substituting from (1), we obtain

$$E\hat{\rho}^m = \int_0^\infty \left[\rho^{-mn/x}\right] \frac{x^{n-1} e^{-x}}{(n-1)!}\, dx \,. \qquad (2)$$

Another estimator of ρ is the MVUE $\tilde{\rho}$

$$\tilde{\rho} = \left(1 - \frac{y_o}{n\hat{\theta}}\right)^{n-1} \left\langle \hat{\theta} > \frac{y_o}{n} \right\rangle ,$$

where $\hat{\theta} = \Sigma y_i/n$. It should be noted that $\tilde{\rho}$ is specifically the MVUE for ρ, and that if one knew with certainty that ρ^m was the parameter of interest, then one should use the MVUE for ρ^m. However, it would appear that Wasan proposes to use $\tilde{\rho}^m$ to estimate ρ^m. We have:

$$E\tilde{\rho}^m = \int_0^\infty \left(1 - \frac{y_o}{\theta x}\right)^{(n-1)m} \left\langle \theta x > y_o \right\rangle \frac{x^{n-1} e^{-x}}{(n-1)!}\, dx$$

$$= \int_0^\infty \left(1 + \frac{\log \rho}{x}\right) \left\langle x > -\log \rho \right\rangle \frac{x^{n-1} e^{-x}}{(n-1)!}\, dx$$

$$= \int_0^\infty \left[\left(1 + \frac{\log \rho}{x}\right) \left\langle 1 + \frac{\log \rho}{x} > 0 \right\rangle\right]^{(n-1)m} \frac{x^{n-1} e^{-x}}{(n-1)!}\, dx. \qquad (3)$$

Neither (2) nor (3) reduce to tabulated functions, so that numerical integration is required.

TRUNCATION ERROR

It may be noted that in both (2) and (3), the term in brackets is always less than one. Thus the truncation error made in carrying out the integration over a finite range $(0,t)$, will be bounded by

$$\int_t^\infty \frac{x^{n-1} e^{-x}}{(n-1)} dx. \qquad (4)$$

For specified values of n and t, (4) may be obtained from a table of the incomplete gamma function. In the program which was used to compute the table, t was set at 22 and n, at 5, so that the truncation error made in evaluating (2) and (3) was less than 0.5×10^{-6}.

INTEGRATION

The integration of (2) and (3) over $(0,22)$ was carried out using Simpson's rule iteratively, with successive halving of the step size. Iterations were terminated when successive approximations of the integral differed by less than 0.25×10^{-6}. Convergence to the integral appeared to be quadratic in all cases; the effective difference between the final approximation and the penultimate one was usually of the order of 10^{-8} or 10^{-7}, and the effective mesh size used to compute the integral was either 0.043 or 0.086 (equivalently, the final approximation involves a sum of 2^{10} or 2^9 functional values).

All the above computations were done in double precision on the IBM 370/158 computer at McGill University. A listing of the program follows this appendix. In the program, $n = 5$ and $\rho = 0.951$, as in Wasan's example, while $m = 1, 2, 10, 20, 40$.

The figures for the mean in the table were taken directly from the computer output; they may be assumed to be correct to the number of decimal places in the table. The root-mean-square error was computed with the aid of two evaluations of $E\hat{\rho}^m$ or $E\tilde{\rho}^m$; e.g.:

$$\text{RMSE}(\hat{\rho}^m) = \sqrt{E(\hat{\rho}^m - \rho^m)^2} = \sqrt{E(\hat{\rho}^{2m}) - 2\rho^m E(\rho^m) + \rho^{2m}}. \qquad (5)$$

The calculation was done an APF Mark 51 scientific hand calculator with arithmetical operations valid to about seven decimal places. However, even though the truncation and approximation errors together may be of order 10^{-6}, the values used in computing (5) had been rounded to 10^{-5} by the program. Therefore, the third decimal in the RMSE figures in the table may not be quite accurate.

```
$WATFIV
C       EXPECTATION OF SOME MOMENTS OF ESTIMATORS OF THE RELIABILITY R
C       FOR THE WEIBULL DISTRIBUTION (LOCATION = 0, SCALE = UNKNOWN,
C       SHAPE = KNOWN ) OR EQUIVALENTLY FOR THE ONE-PARAMETER EXPONENTIAL
C
C        PROGRAMMED 1978.04.27 BY ALEX MCMILLAN
C
C       TO OBTAIN A RUN FOR THE M.L.E., PLACE A 'C' IN COLUMN 1 OF THE
C       FIRST CARD IN EACH OF THE TWO PAIRS LABELED 50 AND 1000 AND REMOVE
C       THE 'C' FROM THE SECOND CARD OF EACH PAIR
C
      1       IMPLICIT REAL*8 (A-H,P-Z)
C       MOMNT IS A TABLE OF THE IN MOMENTS TO BE COMPUTED
      2       INTEGER MOMNT(5) / 1,2,10,20,40 / , IN /5/
C
      3       DDIM(A,B) = A - DMIN1(A,B)
C       F IS A STATEMENT FUNCTION CORRESPONDING TO THE UNNORMALIZED
C       INTEGRAND
      4    50 F(X) = DDIM( 1.0D0 + DLOGR/X , 0.0D0 )**NS * X**M * DEXP(-X)
C     4  50 F(X) = R**(NS/X) * X**M * DEXP(-X)
C       RANGE = ZERO TO HI, TRUNCATION ERROR BOUND PROP(GAMMA(N) .GT. HI)
      5       HI = 22
C       R = RELIABILITY
      6       R = 0.95D0
      7       DLOGR = DLOG(R)
C       N = NUMBER OF ITEMS IN THE SAMPLE USED TO ESTIMATE R
      8       N = 5
      9       M = N - 1
C       STOPPING RULE = SUCCESSIVE INTEGRALS DIFFERING BY LESS THAN STOP,
C       OR NITER  ITERATIONS.  XNORM IS A NORMALIZING CONSTANT USED TO
C       REDUCE THE NUMBER OF OPERATIONS IN THE INTEGRAND
     10       NITER = 20
     11       STOP = 0.25D-6
     12       XNORM = 1.0D0 / ( 6.0D0 * DGAMMA( DFLOAT(N) ) )
C       LOOP TO COMPUTE DIFFERENT MOMENTS = EXPECTATION OF ESTIMATED
C       RELIABILITY OF A SYSTEM WITH 'IS' COMPONENTS
     13       DO 1 IMS = 1,IN
     14       IS = MOMNT(IMS)
     15       MS = M * IS
     16       NS = N * IS
C       USE SIMPSON'S RULE SETTING F(0) = 0 TO AVOID A SINGULARITY
     17       ENDS = F(HI)
C       STEP IS TWICE THE ACTUAL STEP SIZE
     18       STEP = HI
     19       SMODD = F( HI/2D0 )
     20       SMEVEN = 0.0
     21       K = 1
     22       SMOLD = (ENDS + 4D0*SMODD) * STEP * XNORM
     23       WRITE(6,1000) IS
     24  1000 FORMAT('1   VALUE OF MOMENT #',I3 ,'   OF M.V.U.E.'/////)
     C1000 FORMAT('1   VALUE OF MOMENT #',I3 ,'   OF M.L.E.  '/////)
     25       WRITE(6,1001) K,SMOLD,SPOLD
     26  1001 FORMAT(' ITER, #',I3,'      VALUE =',E12.5,'    CHANGE =',E12.5)
     27       DO 2 ITER = 2,NITER
     28       SMEVEN = SMEVEN + SMODD
     29       SMODD = 0.0
     30       STEP = STEP/2
     31       K = K * 2
     32       X = HI + STEP/2.0
C       BACKWARD SUMMATION STARTS WITH SMALLER TERMS

     33       DO 3 J = 1,K
     34       X = X - STEP
     35     3 SMODD = SMODD + F(X)
     36       SMNEW = (ENDS + 4D0*SMODD + 2D0*SMEVEN) * STEP * XNORM
     37       DIFF = SMNEW - SMOLD
     38       WRITE(6,1001) ITER,SMNEW,DIFF
     39       SMOLD = SMNEW
     40       IF(DABS(DIFF) .LT. STOP ) GO TO 4
     41     2 CONTINUE
     42     4 RIS = R**IS
     43       ABSER = SMNEW - RIS
     44       RELER = ABSER / RIS
     45       WRITE(6,1002) IS,SMNEW,IS,RIS,ABSER,RELER
     46  1002 FORMAT('0   END OF ITERATIONS, FINAL VALUE OF MOMENT #',I3,'  IS',
         1   F8.5 /'0',T29,' CORRESPONDING R**',I2,'  IS',F8.5
         2   /'0  ABSOLUTE ERROR =',F8.5,'    PERCENT ERROR =',
         3   2PF8.3,/)
     47       1 CONTINUE
     48       WRITE(6,1003)
     49  1003 FORMAT('1 END OF JOB')
     50       STOP
     51       END

$DATA
```

FIGURE A.1

Listing of the program to compute the table in Subsection 2.1.3.

APPENDIX B

UNIMODALITY OF THE LIKELIHOOD FUNCTION FROM THE TWO-PARAMETER CAUCHY DISTRIBUTION

In this appendix we illustrate the use of the criterion of Mäkeläinen, Schmidt and Styan (1976) to prove the unimodality of the likelihood function in a model which appears to have been inadequately treated in the literature.

Copas (1975) considers a model with identically and independently distributed observations from a two-parameter Cauchy distribution. With μ denoting the mode of the distribution and σ denoting the standard scale parameter, the likelihood function is

$$p(x;\ (\mu,\sigma)) = \frac{1}{\pi^n}\frac{1}{\sigma^n}\prod_{i=1}^{n}\left[1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right]^{-1}. \qquad (1)$$

The parameter space is $\{(\mu,\sigma):\ -\infty < \mu < \infty,\ 0 < \sigma < \infty\}$.

The argument used by Copas is as follows: for a fixed $\mu = \mu_o$, $p(x;(\mu_o,\sigma))$ is shown to be unimodal as a function of σ*. It is also shown that whenever $\nabla \log p = 0$, the matrix of second derivatives $\nabla\nabla^T \log p$ is negative definite, so that local minima or saddle points are excluded. From these, it is deduced that $p(x;\ (\cdot,\ \cdot))$ must be unimodal (except for samples where at least half of the observations coincide). The above argument is weak on two grounds. First, it fails to take account of the behaviour of the function on the boundary. Second, there exist functions which satisfy both conditions but which nevertheless fail to be unimodal.

* Copas' usage of the term 'unimodal' includes the term 'strictly monotone'.

For example, the function exhibited by Tarone & Gruenhage (1975):

$$f(x,y) = -(e^{-2y} + e^{-y}\sin x) \quad (-\infty < x < \infty\ ,\ -\infty < y < \infty)$$

has local maxima at all points $(x,y) = (\frac{3}{2}\pi + 2\pi k,\ \log 2)$ for any integer k, and is also unimodal in one coordinate: for fixed $x = x_o$, $f(x_o,y)$ is unimodal when considered as a function of y.

It seems rather rare for the likelihood function of an identifiable parameter to be periodic, so f might not be considered to be a convincing counterexample; also, f cannot be related in any obvious way to an actual likelihood function. It may, therefore, be useful to present a second counterexample which looks more like a likelihood function.

Consider first the function

$$g(x,y) = \tfrac{1}{2}[\varphi(x - 1.5) + \varphi(x + 1.5)]\ \varphi(y - x^2)\ ,$$

where $\varphi(\cdot)$ is the density of the standard Gaussian distribution. Then g has only three critical points, at $(x,y) = (0,0)$, (a,a^2) and $(-a,a^2)$, where $a = 1.4632...$ is one of the two modes of $\varphi(x - 1.5) + \varphi(x + 1.5)$. The point $(0,0)$ is a saddle point while (a,a^2) and $(-a,a^2)$ are both local maxima. Hence by restricting the parameter space

$$h(x,y) = g(x,y) \quad (y > 1)\ ,$$

h has two maxima, both in the interior of the carrier set, and no other critical points. Also, for fixed $x = x_o$, $h(x_o,y)$ is unimodal, with mode at $y = x_o^2$ $(|x| > 1)$, $+ 1$ $(|x| \leq 1)$.

The function h could be interpreted as the likelihood function of a bivariate observation $(0,0)$, where the first component is assumed to be a realization of the mixture of a Gaussian distribution with mean at $x + 1.5$ and of a Gaussian distribution with mean at $x - 1.5$, while the

second component is assumed to be independent of the first and distributed as a Gaussian variate with mean at $y - x^2$ (the standard errors of all three Gaussian distributions being known to be equal to one).

Returning now to the Cauchy example, we can see that we need only show constancy on the boundary in order to complete Copas' proof of the unimodality of the likelihood function. For this it is necessary to assume (as does Copas), that one half or more of the observations are not all equal.

Let the distinct values of the x's be denoted $y_{(1)}, y_{(2)}, \ldots, y_{(m)}$, with $y_{(1)} < y_{(2)} < \ldots < y_{(m)}$, let n_i be the number of x's equal to $y_{(i)}$ and $n_0 = \max_i n_i$; finally denote $d = \frac{1}{2}\min_i (y_{(i)} - y_{(i-1)})$. Under the above assumption, $n \geq 2n_0$ or, equivalently, $n - 2n_0 \geq 1$. Our approach is to consider the value of the likelihood function $p(x; (\mu,\sigma))$ on a rectangle $R(N)$ in the parameter space with vertices at $(\mu,\sigma) = (\pm N, 1/N)$, $(\pm N, N)$. As N increases, the rectangle clearly approaches the boundary of the parameter space, so that we can restrict our attention to

$$N > \max\{-3y_{(1)}, 3y_{(n)}, 1\} . \tag{2}$$

It is easily seen that with N as in (2), we have for all i :

$$|y_{(i)} - N| < \tfrac{2}{3} N \tag{3}$$

$$\text{and } |y_{(i)} + N| < \tfrac{2}{3} N .$$

Clearly $R(N) = R_1 \cup R_2 \cup R_3 \cup R_4$, where

$$R_1 = \{(\mu,\sigma): -N \leq \mu \leq N, \ \sigma = 1/N\}$$

$$R_2 = \{(\mu,\sigma): -N \leq \mu \leq N, \ \sigma = N\}$$

$$R_3 = \{(\mu,\sigma): \mu = N, \ 1/N \leq \sigma \leq N\}$$

$$R_4 = \{(\mu,\sigma): \mu = -N, \ 1/N \leq \sigma \leq N\} .$$

First note that there is at most one i such that $|y_{(i)} - \mu| < d$.

For all other $j \neq i$, therefore, $|y_{(j)} - \mu| > d$, which ensures

$$\left[1 + \left(\frac{y_{(j)} - \mu}{\sigma}\right)^2\right]^{-1} \leq \left[1 + \frac{d^2}{\sigma^2}\right]^{-1},$$

while

$$\left[1 + \left(\frac{y_{(i)} - \mu}{\sigma}\right)^2\right]^{-1} \leq 1.$$

This ensures that

$$\pi^n p(x;\ (\mu,\sigma)) \leq \frac{1}{\sigma^n}\left[1 + \frac{d^2}{\sigma^2}\right]^{-(n-n_0)}$$
$$\leq \sigma^{n-2n_0}\, d^{-2(n-n_0)}.$$

The last relation enables us to get a crude bound for p on R_1:

$$\pi^n p(x;\ (\mu,\sigma)) \leq c_1 \frac{1}{N},$$

where $c_1 = d^{-2(n-n_0)}$.

Turning now to R_2, we see immediately that

$$\pi^n p(x;\ (\mu,\sigma)) \leq \frac{1}{\sigma^n} = \frac{1}{N^n} < \frac{1}{N}.$$

On R_3, (3) leads to

$$\sigma + \left(\frac{y_{(i)} - N}{\sigma}\right)^2 \geq \sigma + \frac{4}{9}\frac{N^2}{\sigma}$$

so that

$$\pi^n p(x;\ (\mu,\sigma)) \leq \sigma^n\left(\sigma^2 + \frac{4}{9}N^2\right)^{-n}$$
$$\leq N^n\left(\frac{4}{9}N^2\right)^{-n}$$
$$\leq c_2 \frac{1}{N^n} \leq c_2 \frac{1}{N},$$

where $c_2 = \left(\frac{9}{4}\right)^n$.

Similarly, on R_4:

$$\pi^n p(x;\ (\mu,\sigma)) \leq c_2 \frac{1}{N}.$$

It follows from the above that on $R(N)$

$$p(x;\ (\mu,\sigma)) \leq \frac{c_3}{\pi^n}\frac{1}{N},$$

where $c_3 = \max(c_1,c_2)$. As $N \to \infty$, $p(x;\ (\mu,\sigma)) \to 0$ on $R(N)$ so that the likelihood function is constant (zero) on the boundary of the parameter space.

The situation where $2n_0 > n$, is discussed in Copas (1975). It should be noted that when so many observations are coincident, the continuous model would seem to be a rather crude approximation to the data, and consideration should be given to computing the exact likelihood function for the discretized observations, as discussed in subsection 4.1.3.

Part of the argument presented in this section is taken from a preliminary revision for publication of Mäkeläinen, Schmidt & Styan (1976).

APPENDIX C

NOTES ON MINIMUM ENTROPY ESTIMATION

Kriz & Talacko (1968) have attempted to develop a 'minimum entropy estimator' using a measure of information defined as the entropy of the posterior distribution minus the entropy of the prior distribution concentrated at the true value of the parameter. Their 'minimum entropy estimator' is the estimator which minimizes the loss of information when the true parameter is replaced by an estimated value. Their claim is that in the case of independently and identically distributed observations the MLE corresponds to their minimum entropy estimator; in the case where the observations are dependent or do not follow the same distribution they indicate that the MLE would not have the stated optimality property.

Unfortunately, we believe the claim made in the above paper to be unfounded in its wide generality and perhaps meaningless. In *Mathematical Reviews*, Körezlioğlu (1969) notes the need for correction at one point but does not pass judgement on the result. Nevertheless, we indicate below some of the major points where we believe the argument is wrong. We have not succeeded in repairing the proof and conjecture that the result does not hold. The article's original notation is used to a great extent. In our notation

ξ_i	would be	x_i ,	the i-th observation in the data
ξ	" "	$\underset{\sim}{x}$,	the data considered as a point in n-dimensional Euclidean space
θ	" "	$\theta \in \Theta$	(used here as a dummy variable)
α	" "	θ_o ,	the true parameter
$f(\xi_i \mid \theta)$	" "	$q(x_i; \theta)$,	the density of the i-th component of $\underset{\sim}{x}$ (properly, the parent distribution)
$H(\xi \mid \theta)$	would be		a vector whose i-th component is $H(q(\cdot; \theta))$, the entropy of $q(\cdot; \theta)$.

(1) On page 58 of the article, in formula '9' the substitution:

$$f(\xi_i \mid \hat{\theta}(\xi)) = \int_{\Theta} f(\xi_i \mid \theta)\ \delta(\theta - \alpha)\ d\theta \qquad (9)$$

is made, where $\delta(\cdot - \alpha)$ is the degenerate prior concentrated at α. Two interpretations may be imagined for ξ on the left-hand side of (9). First, that ξ is a dummy variable representing a general sample point which is the integration variable; this interpretation is belied by the fact that $\hat{\theta}$ for a given sample is not exactly equal to the true parameter value α. The second interpretation would have ξ be the collection (in vector form) of the sample points ξ_i. But then the left-hand side of (9) would depend strictly on the observed data, while the right-hand side would depend on both the data and the parameter α, which would mean that the parameter α does not index the distribution. The objection might be circumvented by requiring $\hat{\theta}$ to be sufficient (which the article does not do) but this would narrow the applicability of the result.

(2) In Section 3, pp.59-50, the quantity $Q(\hat{\theta};\ \xi) = H(\xi|\hat{\theta}) - H(\xi|\alpha)$ is the difference in the loss of information by taking the parameter to be its estimated value rather than the true value. According to the definition, the norm of Q should be minimized; instead the norm of $H(\xi|\hat{\theta})$ is minimized. Even though Q depends on the choice of estimator only through $H(\xi|\hat{\theta})$, the estimator which minimizes the norm of $H(\xi|\hat{\theta})$ will not necessarily be the estimator which minimizes the norm of Q.

(3) Again on page 60, the quantity

$$[\ \Sigma_i(-\ \log^2 f(\xi_i|\hat{\theta})\ f^2(\xi_i|\alpha))]^{\frac{1}{2}} \qquad (10)$$

would appear to be a misprint for

$$|\ \Sigma(\log f(\xi_i|\hat{\theta}))^2\ f^2(\xi_i|\alpha)\ ]^{\frac{1}{2}}\ , \qquad (11)$$

since (10) is an imaginary number, But the transformation of (11) to:

$$\Sigma\ -\ \log f(\xi_i|\hat{\theta})\ f(\xi_i|\alpha)$$

is puzzling, since in general $[\ \sum_i^n y_i^2\ ]^{\frac{1}{2}} \neq \sum_i^n y_i$.

Since the above steps appear to be central to the argument, the result concerning the optimality property of the MLE is in doubt.