جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل/ كلية التربية

حول الخواص التقاربية لمقدرات الأمكان الأعظم

بحث
مقدم الى قسم الرياضيات- كلية التربية- جامعة بابل وهو جزء من متطلبات
نيل درجة الماجستير في علوم الرياضيات

من قبل

رواسي عدنان حميد المسعودي

بإشراف

الاستاذ المساعد الدكتور سعد عبد ماضي

1424هـ 2003 م

Republic of Iraq,
Ministry of Higher Education,
and Scientific Research,
Babylon University,
College of Education

ON ASYMPTOTIC PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

A Research
Submitted to the Department of Mathematics- College of Education-
University of Babylon in Partial fulfillment of the requirements
for the Degree of Master of Science
in Mathematics

By
Rawassy A. Hameed Al-Masa′oodi

Supervised by
Assist. Prof. Dr. Sa′ad A. Madhi

2003

ABSTRACT

A central role is played in large sample parametric inference by the maximum likelihood estimation of the parameter of interest. This project deals with aspects of large sample behavior of the maximum likelihood estimators such as consistency and asymptotic normality.

For rigorous and complete proofs, analytic details and regularity conditions are introduced.

Illustrative examples and some concluding remarks are also given.

حول الخواص التقاربية لمقدرات الأمكان الأعظم

الخلاصة

يلعب مقدر الامكان الاعظم للمعلمة مثار الاهتمام دورا مركزياً في الاستدلال المعلمي في حالة العينة الكبيرة.

هذا البحث يتناول مظاهر سلوك مقدر الامكان الاعظم في حالة العينة الكبيرة، مثل الاتساق وسلوك التوزيع الطبيعي تقريباً وغيرها.

ولكي تكون البراهين كاملة ودقيقة تم عرض التفاصيل التحليلية المطلوبة مع الافتراضات المناسبة لها.

وتضمن البحث ايضاً بعض الامثلة التوضيحية للنتائج بالاضافة الى بعض الملاحظات حول هذه النتائج.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest thanks for God without whose support and mercy this work would not have been finished.

I also extend my thanks to my supervisor, Assist. Prof. Dr. Sa’ad Abid Madhi, for his consitructive remarks, advice, and comments. I’m also acknowledge a great deal of gratitude for Assist. Prof. Dr. Lua’y Abdul-Hani Al- Suwidi for his continual encouragement and support.

I feel happy to express my thanks and gratitude for the staff of the Department of Mathematics/ College of Education/ University of Babylon for their cooperation.

Also I thank all those who helped me very much in the fulfilment of this work.

Endless thanks are due to my family for their spiritual support during the period of writing this research.

الاهداء

اهدي بحثي هذا الى من شجعني دائماً واحاطني بالعطف والحنان ابداً وقام بمساعدتي قبل واثناء اتمامي لبحثي هذا وهو عائلتي (ابي وامي واخوتي واختي) شيء بسيط لما قاموا به لاجلي.

وهو ايضاً مهداة الى اقربائي عائلة السيد عبود السيد مهدي الشلاه بجميع افرادها لدعمهم اللامحدود لي ومساعدتهم المستمرة لي.

ومن الله التوفيق

CONTENTS

SUPERVISOR CERTIFICATION

I certify that this research was prepared under my supervision at the Department of Mathematics/ College of Education/ University of Babylon, as a partial fulfilment of the requirements for the degree of Master of Science in Mathematics.

Signature:

Name: ***Assist. prof.DR. SA'AD A. MADHI***

Date:

RECOMMENDATION OF THE HEAD OF THE DEPARTMENT

In view of the available recommendations, I forward this research for debate by the Examining Committee.

Signature:

Name: Assist. Prof. Dr. Lua'y Abdul Hani Al-Siweedy

Date:

CERTIFICATION

We certify that we have read this research entitled “ On Asymptotic Properties of Maximum Likelihood Estimator” and, as an Examining Committee, we examined the student in the contents, and what is connected with, and that in our opinion it meets the standard of a research for the degree of Master of Science in Mathematics.

Chairman

Signiture:

Name:

Date:

Member

Signiture:

Name:

Date:

Member

Signiture:

Name:

Date:

Member (Supervisor)

Signiture:

Name:

Date:

Approved by the Dean of the College of Education

Signature:

Name:

Date:

بسم الله الرحمن الرحيم

﴿وقل اعملوا فسيرى الله عملكم

ورسوله والمؤمنون﴾

صدق الله العظيم

سورة التوبة
الآية (105)

INTRODUCTION

The chief justification for the method of maximum likelihood estimation lies in its near-optimal properties for large samples.

This project deals with the asymptotic properties of the maximum likelihood estimators such as consistency and asymptotic normality. To do this rigorously requires analytic details and suitable regularity conditions. Some examples are presented to illustrate the results. Concluding remarks are also given.

A summary of the chapters is as follows:

Chapter two provides some concepts and results that are needed in later chapters.

The large –sample properties of the maximum likelihood estimators are studied in chapter three.

Chapter four contains examples illustrating the results in chapter three.

Finally, in chapter five some comments on the large–sample properties of the maximum likelihood estimator are presented.

CHAPTER One

PRELIMINARIES

In this chapter, we cover thoroughly many concepts and results which find use in later chapters. We refer back to this chapter if and as the need arise.

1.1- Definitions and results from Analysis and Topology

1.1.1- Limit of sequences [1]

Let {A$_n$} be a sequence of sets. The set of all points ω in the set Ω that belong to A$_n$ for infinitely many values of n is called the limit superior of the sequence {A$_n$} and is written as $\limsup\limits_{n\to\infty}\ \mathrm{A}_n$.

The set of all points that belong to A$_n$ for all but a finite number of values of n is called the limit inferior of the sequence {A$_n$} and is written as $\liminf\limits_{n\to\infty}\mathrm{A}_n$.

If $\limsup\limits_{n\to\infty}\ \mathrm{A}_n = \liminf\limits_{n\to\infty}\ \mathrm{A}_n$,
we say that limit exists of the sequence {A$_n$} and we denote it by $\lim\limits_{n\to\infty}\ \mathrm{A}_n$.

For an increasing sequence $A_1 \subset A_2 \subset ...,$ $\lim\limits_{n\to\infty}\mathrm{A}_n = \bigcup\limits_{n}\mathrm{A}_n$.

For a decreasing sequence $A_1 \supset A_2 \supset ...,$ $\lim\limits_{n\to\infty}\mathrm{A}_n = \bigcap\limits_{n}\mathrm{A}_n$.

In the case of an arbitrary sequence of sets A$_1$, A$_2$,… , we have $\liminf\limits_{n\to\infty}\mathrm{A}_n = \bigcup\limits_{n=1}^{\infty}\bigcap\limits_{k=n}^{\infty}\mathrm{A}_\mathrm{k}$, $\limsup\limits_{n\to\infty}\mathrm{A}_n = \bigcap\limits_{n=1}^{\infty}\bigcup\limits_{k=n}^{\infty}\mathrm{A}_\mathrm{k}$.

1.1.2- Upper Semi Continuous Function [4]

A real–valued function $f(\underline{\theta})$ defined on the set Θ is said to be upper semi continuous on Θ, if for all $\underline{\theta}$ in Θ and for any sequence $\underline{\theta}_n$ in Θ, such that $\underline{\theta}_n \to \underline{\theta}$, we have

$$\limsup_{n\to\infty} f(\underline{\theta}_n) \le f(\underline{\theta}\ )$$

or, equivalently, if for all $\underline{\theta}$ in Θ,

$$\sup_{|\underline{\theta}'-\underline{\theta}|<\rho} \quad f(\underline{\theta}') \to f(\underline{\theta}) \quad \text{as} \quad \rho \to 0\ .$$

1.1.3- Taylor Expansion [4]

If $f : R^k \to R$ and if $f''(\underline{x})$ is continuous in the sphere $\{\underline{x} :| \underline{x} - \underline{x}_\circ | < r\}$, then for $|\underline{t}| < r$,

$$f(\underline{x}_\circ + \underline{t}) = f(\underline{x}_\circ) + f'(\underline{x}_\circ)\underline{t} + \underline{t}^T \int_0^1\!\!\int_0^1 \lambda \ \ f''(\ \underline{x}_\circ + u\lambda\,\underline{t})dud\lambda\,\underline{t}$$

where R^k is the k-dimensional Euclidean space.

1.1.4- Convex Function [3]

Let $U \subseteq R^d$ and $\underline{x}_1$ and $\underline{x}_2 \in R^d$, then the point $\underline{x} = (1-\lambda)\underline{x}_1 + \lambda\,\underline{x}_2,\ \ 0 \le \lambda \le 1,$ is called the convex linear combination of $\underline{x}_1$ and $\underline{x}_2$. It is any point lying on the line segment joining $\underline{x}_1$ and $\underline{x}_2$. A set $U \subseteq R^d$ is said to be convex, if the convex linear combination of any two points in U belongs to U .

In other words, U is convex if $\underline{x}_1, \underline{x}_2 \in U$, implies that $\underline{x} \in U$, where $\underline{x} = (1-\lambda)\underline{x}_1 + \lambda\,\underline{x}_2,\ \ 0 \le \lambda \le 1$.

If $\underline{x} \in U \subset R^d$, where U is a convex set, then the function $f(\underline{x})$ is said to be convex if for any two points $\underline{x}_1$ and $\underline{x}_2$ in U ,

$$f(\underline{x}) \le (1-\lambda)\ f(\underline{x}_1) + \lambda\ f(\underline{x}_2) \qquad\qquad ,\ 0 \le \lambda \le 1$$

for every $\underline{x} = (1-\lambda)\underline{x}_1 + \lambda\,\underline{x}_2$.

The function is said to be concave if the inequality sign is reversed or if - $f(\underline{x})$ is convex. If the inequality is strict, then f is called strictly convex function.

1.1.5- σ-algebra (σ-field) [1]

A non-empty class *F* of subsets of a set Ω which is closed under the formulation of countable unions and complements and contains Ø is known as a σ-field.

1.1.6- Measure [1]

A non- negative extended real-valued set function μ defined on a class *F* of subsets of a set Ω is called a measure on *F*, if μ(A) ≥ 0, for all $A \in F$, and F is $\sigma -$ algebra, and $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.

1.1.7- Measurable Space [1]

A measurable space is a pair (Ω, *F*) when Ω is a set and *F* is a σ-field of subsets of Ω. Any member of *F* is called a measurable set.

1.1.8- Measurable Function [1]

Let (Ω, *F*) and (Ω', *F* ') be two measurable spaces.

A function $f: \Omega \to \Omega'$ is said to be measurable (relative to *F* and *F* ') if $f^{-1}(A) \in F$, for every $A \in F'$.

1.1.9- σ-Finite Measure [1]

An extended real valued set function *v* defined on a class *F* of subset of a set Ω is said to be σ-finite measure if for each $A \in F$, there is a sequence of sets $A_n \in F$, such that

$A \subset \bigcup_n A_n$, and $v\,(A_n)$ is finite for all *n*.

i.e $v\,(A_n) < \infty$.

1.1.10- Topological Space [7]

Let Ω be a non-empty set. A class *F* of subsets of Ω is a topology on Ω iff *F* satisfies the following axioms:

1- Ω and Ø belong to *F*.

2- The union of any number of sets in *F* belongs to *F*.

3- The intersection of any two sets in *F* belongs to *F*. The members of *F* are then called *F*-open sets, or simply open sets. The pair (Ω, *F*) is called a topological space.

1.1.11- Open Cover [7]

Let $\mathrm{C} = \{\mathrm{G}_\alpha : \alpha \in \wedge\}$ be open collection of subsets of a topological space X and a subset A of X, we say that C is an open cover of the space X if $\mathrm{X} = \cup\{\mathrm{G}_\alpha : \alpha \in \wedge\}$.

1.1.12- Finite Sub-Cover [7]

A subcollection C' of C , such that C' covers A. An open cover C of A is said to be finite sub-cover if it consists of finite number of open sets C' .v

1.1.13- Compact Set [7]

A subset A of a topological space X is said to be compact iff every open cover of A has a finite sub-cover. More explicitly, that is iff for every collection $\{G_\alpha : \alpha \in \wedge\}$ of open sets for which $A \subset \cup\{G_\alpha : \alpha \in \wedge\}$, there exists finitely many sets $G_{\alpha_1}, G_{\alpha_2}, \ldots, G_{\alpha_n}$ among the Gα's such that:

$$A \subset \cup G_{\alpha_1} \cup \ldots \cup G_{\alpha_n}.$$

1.2- Definitions and results from Probability and Statistics

1.2.1-Modes of Convergence [3], [4]

For a random vector $\underline{x} = (x_1, x_2, \ldots, x_d) \in R^d$, the distribution function of $\underline{x}$, defined for $\underline{x} = (x_1, x_2, \ldots, x_d) \in R^d$, is denoted by:

$$F_{\underline{x}}(\underline{x}) = p(\underline{x} \le \underline{x}) = p(x_1 \le x_1, \ldots, x_d \le x_d) \in R^d.$$

The Euclidean norm of $\underline{x} = (x_1, x_2, \ldots, x_d) \in R^d$ is denoted by $|\underline{x}| = (x_1^2 + x_2^2 + \ldots + x_d^2)^{\frac{1}{2}}$.

Let $\underline{x}$, $\underline{x}_1, \underline{x}_2$, …be random vectors with values in R^d. we have the following modes of convergence.

(a) Convergence in Distribution

The sequence $\{\underline{x}_n\}$ converges in law to $\underline{x}$, $\underline{x}_n \xrightarrow{L} \underline{x}$, if $F_{\underline{x}_n}(\underline{x}) \to F_{\underline{x}}(\underline{x})$ as $n \to \infty$, for all points $\underline{x}$ at which $F_{\underline{x}}(\underline{x})$ is continuous. It is sometimes called convergence in distribution or weak convergence.

(b) Convergence in Probability

The sequence $\{\underline{x}_n\}$ converges in probability to $\underline{x}$, $\underline{x}_n \xrightarrow{p} \underline{x}$ if for every ε >0,

$$p\{|\underline{x}_n - \underline{x}\ | > \varepsilon\} \to 0, \quad \text{as} \quad n \to \infty.$$

(c) Convergence in rth mean

For a real number *r*>0, the sequence $\{\underline{x}_n\}$ converges in the r^{th} mean to $\underline{x}$, $\underline{x}_n \xrightarrow{r} \underline{x}$, if

$$E\,|\underline{x}_n - \underline{x}\ |^r \to 0, \quad \text{as} \quad n \to \infty.$$

where *E* is the expectation value.

(d) Convergence almost surely

The sequence $\{\underline{x}_n\}$ converges almost surely to $\underline{x}$, $\underline{x}_n \xrightarrow{a.s} \underline{x}$, if $\quad p\{\lim\limits_{n\to\infty} \underline{x}_n = \underline{x}\ \} = 1.$

Almost sure converges is sometimes called convergence with probability 1 (w.p.1) or strong convergence.

Equivalently, $\underline{x}_n \xrightarrow{a.s} \underline{x}$ iff for every ε >0,

$$p\{|\underline{x}_m - \underline{x}\ | < \varepsilon, \text{for all} \quad m \geq n\} \to 1 \quad \text{as} \quad n \to \infty.$$

1.2.2-Fatou-Lebesgue Theorem [1], [4]

If $\underline{x}_n \xrightarrow{a.s} \underline{x}$ and if for all *n*, $\underline{x}_n \geq \underline{y}$, for some random vector $\underline{y}$ with $E\,|\underline{y}\,| < \infty$, then $\liminf\limits_{n\to\infty} E(\underline{x}_n) \geq E(\underline{x}\ )$.

1.2.3- Montone Convergence Theorem [4]

If $0 \le \underline{x}_1 \le \underline{x}_2 \le \ldots$ and $\underline{x}_n \xrightarrow{a.s} \underline{x}$, then $E(\underline{x}_n) \to E(\underline{x})$.

1.2.4- Lebesgue Dominated Convergence Theorem [4]

If $\underline{x}_n \xrightarrow{a.s} \underline{x}$ and if $|\underline{x}_n| \le \underline{y}$, for some random vector $\underline{y}$ with $E|\underline{y}| < \infty$, then $E(\underline{x}_n) \to E(\underline{x})$, such that $E(\underline{x})$ exists. In these theorems, $\underline{x}$, $E(\underline{x}_n)$ and $E(\underline{x})$ may take the value $+\infty$.

1.2.5- The Law of Large Numbers [6]

Let $\underline{x}_1, \underline{x}_2, \ldots$ be i.i.d (independent identically distributed) random vectors, and let

$$\overline{\underline{x}}_n = \left(\frac{1}{n}\right)\sum_{j=1}^{n} \underline{x}_j.$$

(a) If $E|\underline{x}| \le \infty$, then $\overline{\underline{x}}_n \xrightarrow{p} \underline{\mu} = E(\underline{x})$. This is the weak law of large numbers (WLLN).

(b) $\overline{\underline{x}}_n \xrightarrow{a.s} \underline{\mu}$ iff $E|\underline{x}| < \infty$ and $\underline{\mu} = E(\underline{x})$. This is the strong law of large number (SLLN).

(c) If $E|\underline{x}|^2 < \infty$, then $\overline{\underline{x}}_n \xrightarrow{qm} \underline{\mu} = E(\underline{x})$

(*qm* means in quadratic mean).

1.2.6- Central Limit Theorem [5]

Let $\underline{x}_1, \underline{x}_2, \ldots,$ be i.i.d random vectors with mean $\underline{\mu}$ and finite covariance matrix, *B*. Then $\sqrt{n}(\overline{\underline{x}}_n - \underline{\mu}) \xrightarrow{L} N(\underline{0}, B)$.

1.2.7- The Continuity Theorem [4]

$$\underline{x}_n \xrightarrow{L} \underline{x} \quad \text{iff} \quad \varphi_{\underline{x}_n}(\underline{t}\,) \to \varphi_{\underline{x}}(\underline{t}\,), \quad \text{for all} \quad \underline{t} \in R^d$$

where $\varphi_{\underline{x}}$ is the characteristic function of $\underline{x}$.

1.2.8- Jensen's Inequality [3]

Let $\underline{x}$ be a random vector with mean $E(\underline{x})$, and let f(.) be convex function, then

$$E[f(\underline{x})] \geq f[E(\underline{x})].$$

1.2.9- Slutsky Theorems [4]

(a) If $\{\underline{x}_n\}$ be a sequence of random vectors, $x_n \in R^d$, $x_n \xrightarrow{L} \underline{x}$ and if $f : R^d \to R^k$ is such that

$p\{\underline{x} \in c(f)\} = 1$, where $c(f)$ is the continuity set of f,

then $f(\underline{x}_n) \xrightarrow{L} f(\underline{x})$.

(b) If $\underline{x}_n \xrightarrow{L} \underline{x}$ and $(\underline{x}_n - \underline{y}_n) \xrightarrow{p} \underline{0}$, then $\underline{y}_n \xrightarrow{L} \underline{x}$.

(c)If $\underline{x}_n \in R^d$, $\underline{y}_n \in R^k$, $\underline{x}_n \xrightarrow{L} \underline{x}$ and $\underline{y}_n \xrightarrow{L} c$,

then $\begin{pmatrix} x_n \\ - \\ y_n \end{pmatrix} \xrightarrow{L} \begin{pmatrix} x \\ - \\ c \end{pmatrix}$.

1.2.10- Consistency in Statistical Estimation [2], [6]

In such problems the underlying probability, $p_{\theta_\circ}$, depends upon a parameter $\underline{\theta}_\circ \in \Theta$ in R^k, where Θ is the parameter space.

Let $\{\underline{\tilde{\theta}}_n\}$ be a sequence of random vectors, considered as estimators of $\underline{\theta}_\circ$, the true value of the parameter.

We say that $\{\underline{\tilde{\theta}}_n\}$ is weakly consistent sequence estimators of $\underline{\theta}_\circ$ if $\underline{\tilde{\theta}}_n \xrightarrow{p} \underline{\theta}_\circ$ for all $\underline{\theta}_\circ \in \Theta$.

where $p = p_{\underline{\theta}_\circ}$ is the “true” probability distribution. This is sometimes called consistency in probability. We may similarly define strong consistency ($\underline{\tilde{\theta}}_n \xrightarrow{a.s} \underline{\theta}_\circ$) or consistency in quadratic mean $(\underline{\tilde{\theta}}_n \xrightarrow{qm} \underline{\theta}_\circ)$, both of which imply (weak consistency).

The weak (strong) law of large numbers states that the sample mean is a weakly (strongly) consistent estimator of the population mean. That is:

If $\underline{\mathrm{x}}_1 , \underline{\mathrm{x}}_2 ,\ldots,\ \underline{\mathrm{x}}_n$ are i.i.d random vectors with common finite $\underline{\mu}$ and $\underline{\bar{\mathrm{x}}}_n = \frac{1}{n}\sum_{i=1}^{n} \underline{\mathrm{x}}_{\mathrm{i}}$,

then $\underline{\bar{\mathrm{x}}}_n \xrightarrow{p} \underline{\mu}$ as $n \to \infty$ (weak consistency),

and $\underline{\bar{\mathrm{x}}}_n \xrightarrow{a.s} \underline{\mu}$ as $n \to \infty$ (strong consistency).

1.2.11-Maximum likelihood Estimator [3]

Let $\mathrm{x}_1 , \mathrm{x}_2 ,\ldots,\ \mathrm{x}_n$ be identically independent random sample with density $f(x/\underline{\theta})$ with respect to some σ-finite measure *v* (usually Lebesgue measure or counting measure) where $\underline{\theta} \in \Theta \subseteq R^k$. The function

$$L_n(\underline{\theta}) = L_n(\underline{\theta}, \underline{x}) = \underset{i=1}{\overset{n}{\Pi}}\ f(x_i, \underline{\theta}) \qquad \ldots\ldots..(1)$$

Considered as a function of $\underline{\theta}$ is called the likelihood function, where the observed values of $\mathrm{x}_1 , \mathrm{x}_2 ,\ldots,\ \mathrm{x}_n$ are $x_1 , x_2 ,\ldots,\ x_n$. The principle of maximum likelihood estimation consists of choosing as an estimator of $\underline{\theta}$ any function $\underline{\hat{\theta}}_n = \underline{\hat{\theta}}_n\ (x_1,\ldots,x_n)$ such that

$$L_n(\underline{\hat{\theta}}_n) = \sup_{\underline{\theta} \in \Theta}\ L_n(\underline{\theta}_n) \ \ldots\ldots.(2)$$

If $\hat{\underline{\theta}}$ satisfying (2) exists, we call it a maximum likelihood estimate (MLE).

It is convenient to work with logarithm of the likelihood function. $\ell_n(\underline{\theta}) = \log L_n(\underline{\theta})$. Since log is monotone function.

$$\ell_n(\hat{\underline{\theta}}_n) = \sup_{\underline{\theta}\in\Theta} \ell_n(\underline{\theta})$$

When $(\frac{\partial}{\partial\underline{\theta}}) f(x,\underline{\theta})$ exists, one can seek the MLE $\hat{\underline{\theta}}_n$ as

$$\frac{\partial L_n(\underline{\theta})}{\partial\underline{\theta}} = \underline{0} \quad \text{and} \quad \frac{\partial^2 L_n(\underline{\theta}_n)}{\partial\underline{\theta}^2} < 0$$

or equivalently

$$\frac{\partial \ell_n(\underline{\theta})}{\partial\underline{\theta}} = \underline{0} \quad \text{and} \quad \frac{\partial^2 \ell_n(\underline{\theta}_n)}{\partial\underline{\theta}^2} < 0 \text{ ……….(3)}$$

Where $\frac{\partial \ell_n(\underline{\theta})}{\partial\underline{\theta}} = \frac{\partial \log L_n(\underline{\theta})}{\partial\underline{\theta}} = \sum_1^n \frac{\partial}{\partial\underline{\theta}} \log f(x_i,\underline{\theta}) = 0$.

Equation (3) is usually referred to as the likelihood equation.

1.2.12- Fisher Information [8]

Fisher's information function plays an important role in the large sample theory of MLE.

Let

$\psi(x,\underline{\theta}) = (\frac{\partial}{\partial\underline{\theta}} \log f(x/\underline{\theta}))^T$, a k dimensional vector and

$\psi'(x,\underline{\theta}) = \frac{\partial^2}{\partial\underline{\theta}^2} \log f(x/\underline{\theta})$, a $k\times k$ matrix.

The Fisher information matrix for a single observation is defined as $B(\underline{\theta}) = E_{\underline{\theta}}(\psi(\mathrm{x},\underline{\theta})\psi(\mathrm{x},\underline{\theta})^T)$ a $k\times k$ matrix.

Assuming that the partial derivative with respect to $\underline{\theta}$ can be passed under the integral sign in

$$\int f(x/\underline{\theta})dv(x)=1 \quad ; \quad \text{we find}$$

$$E_{\underline{\theta}}(\psi(\mathrm{x},\underline{\theta}))=\int\frac{\partial/\partial\underline{\theta}\; f(x/\underline{\theta})}{f(x/\underline{\theta})}f(x/\underline{\theta})dv(x)$$

$$=\int\frac{\partial}{\partial\underline{\theta}}f(x/\underline{\theta})dv(x)=\underline{0}$$

so that $B(\underline{\theta})$ is in fact the covariance matrix of ψ,

$$B(\underline{\theta})=\mathrm{var}_{\underline{\theta}}(\psi(\mathrm{x},\underline{\theta}))$$

If the second partial derivatives with respect to $\underline{\theta}$ can be passed under the intgeral sign, then

$$\int\left(\frac{\partial^2}{\partial\underline{\theta}^2}\right)f(x/\underline{\theta})dv(x)=\underline{0}, \text{ and}$$

$$E_{\underline{\theta}}(\psi'(\mathrm{x},\underline{\theta}))=\int\left[\frac{\partial}{\partial\underline{\theta}}\frac{(\partial/\partial\underline{\theta})f(x/\underline{\theta})}{f(x/\underline{\theta})}\right]f(x/\underline{\theta})dv(x)$$

$$=\int\frac{f(x/\underline{\theta})(\partial^2/\partial\underline{\theta}^2)f(x/\underline{\theta})-((\partial/\partial\underline{\theta})f(x/\underline{\theta}))^T(\partial/\partial\underline{\theta})f(x/\underline{\theta})}{f(x/\underline{\theta})^2}\times$$

$$f(x/\underline{\theta})dv(x)$$

$$=\underline{0}-\int\psi(x,\underline{\theta})\psi(x,\underline{\theta})^T f(\mathrm{x}/\underline{\theta})dv(x).$$

Thus

$$B(\underline{\theta})=-E_{\underline{\theta}}(\psi'(\mathrm{x},\underline{\theta})).$$

Chapter Two

ASYMPTOTIC PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

The main justification of the method of maximum likelihood is a "large-sample" one, which shows that when an observation provides lots of information about an unknown parameter, the method utilizes essentially all of this information. In this chapter, we study the large-sample (asymptotic) properties of the MLE.

2.1- Consistency of MLE

A key property of the maximum likelihood estimators is their consistency under certain assumptions. The proof of consistency of the MLE that we will present in theorem (3.1.1) is mainly based on the ideas given in Wald (1949), Wolfowiz (1949). We start with the following lemma.

Let $f_0(x)$ and $f_1(x)$ be densities with respect to a $\sigma-$finite measure v.

The kullback- Leibler information number is defined as

$$K(f_0,\, f_1)=E_0\left(\log\frac{f_0(\text{x})}{f_1(\text{x})}\right)=\int(\log\frac{f_0(x)}{f_1(x)})\,f_0(x)dv(x).$$

In this expression, $\log\frac{f_0(x)}{f_1(x)}$ is defined as $+\infty$ if $f_1(x)=0$ and $f_0(x)>0$,so the expectation could be $+\infty$. Although $\log\frac{f_0(x)}{f_1(x)}$ is defined as $-\infty$ when $f_1(x)>0$ and $f_0(x)=0$, the integrand, $\log\left(\frac{f_0(x)}{f_1(x)}\right)f_0(x)$, is defined as zero in this case. $K(f_0, f_1)$ is a measure of the ability of the likelihood ratio to distinguish between f_0 and f_1 when f_0 is true.

<u>Lemma</u> (Shannon-kolmogorov Information Inequality)

Let $f_0(x)$ and $f_1(x)$ be densities with respect to v. Then

$$K(f_0, f_1) = E_0\left(\log\frac{f_0(\mathrm{x})}{f_1(\mathrm{x})}\right) = \int \log\frac{f_0(x)}{f_1(x)}\, f_0(x)dv(x) \geq 0$$

With equality if and only if $f_1(x) = f_0(x)$ (a.e dv).

<u>Proof:</u>

Since log x is strictly convex. Jensen's inequality implies

$$-K(f_0, f_1) = E_0\left(\log\frac{f_1(\mathrm{x})}{f_0(\mathrm{x})}\right) \leq \log E_0\, \frac{f_1(\mathrm{x})}{f_0(\mathrm{x})} \quad ,\ldots\ldots\ldots.. (1)$$

With equality if and only if $\frac{f_1(\mathrm{x})}{f_0(\mathrm{x})}$ is a constant with probability 1 when x has density f_0. That is

if $\frac{f_1(x)}{f_0(x)} = c$, then $E(\log c) = \log c$.

But

$$E_0\, \frac{f_1(\mathrm{x})}{f_0(\mathrm{x})} = \int \frac{f_1(x)}{f_0(x)} f_0(x)dv(x) = \int_{s_0} f_1(x)dv(x) \leq 1 ,\ldots\ldots. (2)$$

Where $S_0 = \{x : f_0(x) > 0\}$, with equality iff S_0 has probability 1 under $f_1(x)$. The combination of (1) and (2) gives

$$-K(f_0, f_1) = \log E\left(\frac{f_1(\mathrm{x})}{f_0(\mathrm{x})}\right) \leq \log 1 = 0$$

or

$K(f_0, f_1) \geq 0$, With equality iff $f_1(x) = f_0(x)$ (a.e. dv).

<u>Theorem (2.1.1)</u>

Let x$_1$, x$_2$,…, x$_n$ be i.i.d with density $f(x/\underline{\theta})$, where $\underline{\theta} \in \Theta$, Θ is the parameter space.

Let $\underline{\theta}_\circ$ denotes the true value of the parameter $\underline{\theta}$. $\{\underline{\hat{\theta}}_n\}$ is the sequence of maximum likelihood estimators of the parameter $\underline{\theta}$.

For the validity of the proof of this theorem. We assume the following assumptions called (regularity conditions):

1- $\Theta \subseteq R^k$ is compact,

2- $f(x/\underline{\theta})$ is upper semi continuous in $\underline{\theta}$ for all x ,

3-there exists a function $K(x)$ such that $E_{\underline{\theta}_\circ} |K(\mathrm{x})| < \infty$, and

$$W(x,\underline{\theta}) = \log \frac{f(x/\underline{\theta})}{f(x/\underline{\theta}_\circ)} \le K(x) \text{ , for all } x \text{ and } \underline{\theta},$$

4-for all $\underline{\theta} \in \Theta$ and for all sufficiently small $\rho > 0$,

$$\sup_{|\underline{\theta}' - \underline{\theta}| < \rho} f(x/\underline{\theta}') \text{ is measurable in } x ,$$

5-(Identifiability), $f(x/\underline{\theta}) = f(x/\underline{\theta}_\circ)$ (a.e dv) $\Rightarrow \underline{\theta} = \underline{\theta}_\circ$,

if these conditions are satisfied, and for each $n \ge 1$ the MLE is unique, then, for any sequence of MLE $\underline{\hat{\theta}}_n$ of $\underline{\theta}$, $\underline{\hat{\theta}}_n \xrightarrow{a.s} \underline{\theta}_\circ$.

i.e. For any $\varepsilon > 0$, and for all $\underline{\theta}_\circ \in \Theta$

$$\lim_{n \to \infty} p\{|\underline{\hat{\theta}}_n - \underline{\theta}_\circ| < \varepsilon\} = 1 \quad \text{holds.}$$

i.e The MLE $\underline{\hat{\theta}}_n$ is consistent.

Proof:

Let $f(x,\underline{\theta},\rho) = \sup_{|\underline{\theta}' - \underline{\theta}| < \rho} W(x,\underline{\theta})$.

Then $f(x,\underline{\theta}_\circ,\rho)$ is measurable in x for all sufficiently small $\rho > 0$ by (4).

By (3), $f(x,\underline{\theta},\rho)$ is bounded above by an integrable function K(x), and $f(x,\underline{\theta_\circ},\rho)\to W(x,\underline{\theta}),\quad$ as $\quad\rho\to 0$ by (2) (since $W(x,\underline{\theta})$ is upper semi continuous in $\underline{\theta}$ for all x .

Therefore, by the Monotone Convergence Theorem, we have

$$\int f(x,\underline{\theta},\rho)\,f(x)dv(x)\to\int W(x,\underline{\theta})\,f(x)dv(x)=\mu(\underline{\theta})\quad\text{as}\quad\rho\to 0\,.$$

Let $\varepsilon>0$. Find $\rho_{\underline{\theta}}$ for each $\underline{\theta}$, so that

$$\int f(x,\underline{\theta},\rho_{\underline{\theta}})\,f(x)dv(x)<\mu(\underline{\theta}\,)+\varepsilon\,.$$

The spheres $S(\underline{\theta},\rho_{\underline{\theta}})=\{\underline{\theta}' :|\,\underline{\theta}'-\underline{\theta}\,|<\rho_{\underline{\theta}}\}$ cover Θ.

Since Θ is compact, then there exists a finite sub-cover, say $\Theta\subset\overset{m}{\underset{1}{\cup}}S(\underline{\theta}_j,\rho_{\underline{\theta}_j})$. For each $\underline{\theta}\in\Theta$ there exists an index j such that $\underline{\theta}\in S(\underline{\theta}_j,\rho_{\underline{\theta}_j})$. From the definition of $f(x,\underline{\theta},\rho)$,

$W(x,\underline{\theta})\le\ f(x,\underline{\theta}_j,\rho_{\underline{\theta}_j})$ for all x . Hence,

$$\frac{1}{n}\sum_{i=1}^{n}W(x_i,\underline{\theta})\le\frac{1}{n}\sum_{1}^{n}f(\mathrm{x}_i,\underline{\theta}_j,\rho_{\underline{\theta}_j})\,.$$

So that

$$\sup_{\underline{\theta}\in\Theta}\frac{1}{n}\sum_{1}^{n}W(x_i,\underline{\theta})\le\sup_{1\le j\le m}\frac{1}{n}\sum_{1}^{n}f(x_i,\underline{\theta}_j\ \rho_{\underline{\theta}_j})\,.$$

Now, apply the SLLN to $\frac{1}{n}\sum_{1}^{n}\ f(x_i,\underline{\theta}_j,\rho_{\underline{\theta}_j})$,

$$p\{\lim_{n\to\infty}\frac{1}{n}\sum_{1}^{n}f(x_i,\underline{\theta}_j,\rho_{\underline{\theta}_j})\le\mu(\underline{\theta}_j)+\varepsilon,\quad\text{for } j=1,\ldots,m\}=1,$$

$$p\{\limsup_{n\to\infty}\ \sup_{1\le j\le m}\frac{1}{n}\sum_{1}^{n}f(x_i,\underline{\theta}_j,\rho_{\underline{\theta}_j})\le\sup_{1\le j\le m}\mu(\underline{\theta}_j)+\varepsilon\,\}=1\,,$$

so, $\quad p\{\limsup_{n\to\infty}\ \sup_{\underline{\theta}\in\Theta}\frac{1}{n}\sum_{1}^{n}W(x_i,\underline{\theta})\le\sup_{\underline{\theta}\in\Theta}\mu(\underline{\theta}\ )+\varepsilon\,\}=1$

since it is true for all $\varepsilon>0$, it is also true for $\varepsilon=0$.

$$\therefore\ p\{\limsup_{n\to\infty}\ \sup_{\underline{\theta}\in\Theta}\frac{1}{n}\sum_{1}^{n}W(x_i,\underline{\theta})\le\sup_{\underline{\theta}\in\Theta}\mu(\underline{\theta})\}=1\,.$$

Now, let $\rho>0$, and $S=\{\underline{\theta}:|\,\underline{\theta}-\underline{\theta_\circ}\,|\ge\rho\}$, then S is compact and

$$p_{\underline{\theta}_\circ}\{\limsup_{n\to\infty} \sup_{\underline{\theta}\in S} \frac{1}{n}\sum_1^n W(x_i,\underline{\theta}) \le \sup_{\underline{\theta}\in S} \mu(\underline{\theta})\} = 1.$$

By Shannon-kolmogorov Information Inequality, we have

$$\mu(\underline{\theta}) = -K(\underline{\theta}_\circ,\underline{\theta}) = \int W(x,\underline{\theta})\, f(x,\underline{\theta}_\circ)dv(x) < 0, \qquad \text{for } \underline{\theta} \in S \,,$$

since $W(\mathrm{x},\underline{\theta})$ is bounded above by an integrable function $K(x)$ and by using the Fatou-Lebesgue Theorem,

$$\begin{aligned}\limsup_{\underline{\theta}'\to\underline{\theta}} \mu(\underline{\theta}') &= \limsup_{\underline{\theta}'\to\underline{\theta}} \mathrm{E}\,(W(\mathrm{x},\underline{\theta}')) \\ &\le E \limsup_{\underline{\theta}'\to\underline{\theta}} W(\mathrm{x},\underline{\theta}') \\ &\le E(W(\mathrm{x},\underline{\theta})) = \mu(\underline{\theta}) \quad \text{for every} \quad \underline{\theta}\in\Theta.\end{aligned}$$

So $\mu(\underline{\theta})$ is upper semi continuous and hence achieves its maximum value on *S*.

Let $\delta = \sup_{\underline{\theta}\in S} \mu(\underline{\theta}), \text{so}\, \delta < 0$ and

$$p_{\underline{\theta}_\circ}\{\limsup_{n\to\infty} \sup_{\underline{\theta}\in S} \frac{1}{n}\sum_1^n W(x_i,\underline{\theta}) \le \delta\} = 1\,.$$

Thus, with probability 1, there exists an N such that for all *n*>N,

$$\sup_{\underline{\theta}\in S} \frac{1}{n}\sum_1^n W(\mathrm{x}_i,\underline{\theta}) \le \delta/2 < 0, \text{say}.$$

But

$$\frac{1}{n}\sum_1^n W(\mathrm{x}_i,\underline{\hat{\theta}}_n) = \sup_{\underline{\theta}\in\Theta} \frac{1}{n}\sum_1^n W(\mathrm{x}_i,\underline{\theta}) \ge 0\,,$$

the sum is equal to 0 for $\underline{\theta} = \underline{\theta}_\circ$.

This implies that $\underline{\hat{\theta}}_n \notin S$ for $n > N$, that is, $|\underline{\hat{\theta}}_n - \underline{\theta}| < \rho$.

Since ρ is arbitrary, we get

$$\lim_{n\to\infty} p\{|\underline{\hat{\theta}}_n - \underline{\theta}| < \varepsilon\} = 1\,.$$

So, the sequence of MLE $\underline{\hat{\theta}}_n$ of $\underline{\theta}$ is strongly consistent.

i.e $\underline{\hat{\theta}}_n \xrightarrow{\text{a.s}} \underline{\theta}_\circ$.

2.2- Asymptotic Normality of MLE

The basis of large-sample tests and confidence intervals is the general property that the MLE has a limiting normal distribution around the true parameter value as mean and with a variance that easily calculated. Indeed, this is a much stronger result than consistency and requires additional regularity conditions.

Theorem (2.2.1)

Let x_1, x_2,… be i.i.d with density $f(x/\underline{\theta})$ (with respect to dv), and let $\underline{\theta}_\circ$ denote the true value of the parameter. We make the following regularity conditions.

1. Θ is an open subset of R^k,
2. second partial derivatives of $f(x/\underline{\theta})$ with respect to $\underline{\theta}$ exist and are continuous for all x, and may be passed under the integral sign in $\int f(x/\underline{\theta})dv(x)$,
3. there exists a function $K(x)$ such that $E_{\underline{\theta}_\circ}(K(\text{x}))<\infty$ and each component of $\psi'(x,\underline{\theta})$ is bounded in absolute value by K(x) uniformly in some neighborhood of $\underline{\theta}_\circ$,
4. $B(\underline{\theta}_\circ)=-E_{\underline{\theta}_\circ}[\psi'(\text{x},\underline{\theta}_\circ)]$ is positive definite, the information matrix for a single observation,
5. $f(x/\underline{\theta})=f(x/\underline{\theta}_\circ)\,(\text{a.e}\,dv) \quad \Rightarrow \underline{\theta}=\underline{\theta}_\circ$,

then, there exists a strongly consistent sequence $\underline{\hat{\theta}}_n$ of roots of the likelihood equation such that

$$\sqrt{n}(\underline{\hat{\theta}}_n-\underline{\theta}_\circ)\xrightarrow{L}N(\underline{0},B(\underline{\theta}_\circ)^{-1}).$$

Proof:

Let $S_\rho = \{\underline{\theta} : |\underline{\theta} - \underline{\theta}_\circ| \le \rho\}$, for some $\rho > 0$, be a compact neighborhood of $\underline{\theta}_\circ$ on which components of $\psi'(x, \underline{\theta})$ are uniformly bounded by *K*(x) as in (3). The existence of a strongly consistent sequence $\underline{\hat{\theta}}_n$ of solutions of $\frac{\partial \ell_n(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$ follows from theorem (2.1.1) with Θ=*S*$_\rho$. Conditions (1), (2) and (5) of that theorem (2.1.1) are already satisfied. Condition (4) follows from continuity of $f(x/\underline{\theta})$ in $\underline{\theta}$.

For checking condition (3), we expand $W(x, \underline{\theta})$ about $\underline{\theta}_\circ$ using Taylor expansion as

$$W(x, \underline{\theta}) = w(x, \underline{\theta}_\circ) + \psi(x, \underline{\theta}_\circ)^T (\underline{\theta} - \underline{\theta}_\circ) + (\underline{\theta} - \underline{\theta}_\circ)^T \int_0^1 \int_0^1 \lambda \psi'(x, \underline{\theta}_\circ + u\lambda(\underline{\theta} - \underline{\theta}_\circ))\, du d\lambda\, (\underline{\theta} - \underline{\theta}_\circ).$$

Since $W(x, \underline{\theta}_\circ) = 0, \psi(x, \underline{\theta}_\circ)$ is integrable,and the components of ψ' are bounded by *K*(x) uniformly on S_ρ, hence $W(x, \underline{\theta})$ is bounded uniformly on S_ρ by an integrable function. We have $D_{\underline{\theta}} \ell_n(\underline{\theta}) = \sum_1^n \psi(\mathrm{x}_i, \underline{\theta})$, where $D_{\underline{\theta}}$ is the differential operator $\frac{\partial}{\partial \underline{\theta}}$.

By Taylor theorem, we expand $D_{\underline{\theta}} \ell_n$ about $\underline{\theta}_\circ$ we get

$$D_{\underline{\theta}} \ell_n(\underline{\theta}) = D_{\underline{\theta}} \ell_n(\underline{\theta}_\circ) + \int_0^1 \sum_1^n \psi'(\mathrm{x}_i, \underline{\theta}_\circ + \lambda(\underline{\theta} - \underline{\theta}_\circ)) d\lambda (\underline{\theta} - \underline{\theta}_\circ)$$

$$D_{\underline{\theta}} \ell_n(\underline{\theta}) = D_{\underline{\theta}} \ell_n(\underline{\theta}_\circ) + \int_0^1 \sum_1^n \psi'(\mathrm{x}_i, \underline{\theta}_\circ + \lambda(\underline{\theta} - \underline{\theta}_\circ)) d\lambda (\underline{\theta} - \underline{\theta}_\circ).$$

Now let $\underline{\theta} = \underline{\hat{\theta}}_n$ where $\underline{\hat{\theta}}_n$ is any strongly consistent sequence satisfying

$$D_{\underline{\theta}} \ell_n(\underline{\hat{\theta}}_n) = \underline{0} \quad \text{and divide by } \sqrt{n}$$

$$\frac{1}{\sqrt{n}} D_{\underline{\theta}} \ell_n(\underline{\theta}_\circ) = A_n \sqrt{n} (\underline{\hat{\theta}}_n - \underline{\theta}_\circ),$$

where

$$A_n = -\int_0^1 \frac{1}{n} \sum_1^n \psi'(\mathrm{x}_i, \underline{\theta}_\circ + \lambda(\underline{\hat{\theta}}_n - \underline{\theta}_\circ)) d\lambda.$$

Because $E_{\underline{\theta}_\circ}(\psi(x, \underline{\theta}_\circ)) = \underline{0}$ and $Var_{\underline{\theta}_\circ}(\psi(x, \underline{\theta}_\circ)) = B(\underline{\theta}_\circ)$, hence by the central limit theorem, we obtain that

$$\frac{1}{\sqrt{n}} D_{\underline{\theta}}\ell_n(\underline{\theta}_\circ) = \sqrt{n}(\frac{1}{n}\sum_{\underline{1}}^{n}\psi(x_i,\underline{\theta}_\circ)) \xrightarrow{L} Z \quad N(\underline{0}, B(\underline{\theta}_\circ)).$$

Now, let $\varepsilon > 0$. Since $E_{\underline{\theta}_\circ}(\psi'(x,\underline{\theta}))$ is continuous in $\underline{\theta}$ from condition (3), so there is a $\rho > 0$ such that

$$|\,\underline{\theta} - \underline{\theta}_\circ\,| < \rho \text{ implies } |\,E_{\underline{\theta}_\circ}(\psi'(\mathrm{x},\underline{\theta})) + B(\underline{\theta}_\circ)\,| < \varepsilon .$$

From the SLLN, that with probability 1 there is an integer N such that *n*>N implies

$$\sup_{\underline{\theta}\in S_\rho} |\,\frac{1}{n}\sum_{1}^{n}\psi'(\mathrm{x}_i,\underline{\theta}) - E_{\underline{\theta}_\circ}(\psi'(\mathrm{x},\underline{\theta}))\,| < \varepsilon .$$

Then assuming N is so large that *n*>N implies

$|\,\hat{\underline{\theta}}_n - \underline{\theta}_\circ\,| < \rho$, *n*>N implies

$$|\,A_n - B(\underline{\theta}\,)\,| \le \int_0^1 |\,\frac{1}{n}\sum_{1}^{n}\psi'(\mathrm{x}_i,\underline{\theta}_\circ + \lambda\,(\hat{\underline{\theta}}_n - \underline{\theta}_\circ)) + B(\underline{\theta}_\circ)\,|\,d\lambda$$

$$\le \int_0^1 \sup_{\underline{\theta}\in S_\rho}[\,|\,\frac{1}{n}\sum_{1}^{n}\psi'(\mathrm{x}_i,\underline{\theta}) - E_{\underline{\theta}_\circ}(\psi'(\mathrm{x},\underline{\theta}))\,| +$$

$$|\,E_{\underline{\theta}_\circ}(\psi'(\mathrm{x},\underline{\theta})) + B(\underline{\theta}_\circ)\,|\,]\,d\lambda \le 2\varepsilon .$$

Hence $A_n \xrightarrow{a.s} B(\underline{\theta}_\circ)$ implies A_n^{-1} will exist and by Slutsky’s theorem

$\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_\circ) = A_n^{-1}\frac{1}{\sqrt{n}} D_{\underline{\theta}}\ell_n(\underline{\theta}_\circ) \xrightarrow{L} B(\underline{\theta}_\circ)^{-1} Z \quad N(\underline{0}, B(\underline{\theta}_\circ)^{-1})$. Hence, under the conditions of theorem (3.2.1) if there is a unique solution of the likelihood equation for every *n*, this sequence of solutions will be consistent and asymptotically normal.

2.3- Asymptotic Efficiency of MLE

Under the conditions of theorem (2.2.1), the MLE $\hat{\underline{\theta}}_n$ was seen to be asymptotically unbiased in a reasonably strong sense, because whatever be the true value of $\underline{\theta}$, $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}\,)$ is asymptotically normal with mean zero. Moreover, the asymptotic variance of the MLE is $\frac{1}{n}B(\underline{\theta})^{-1}$ which is the

Crame'r-Rao lower bound for the variance of any unbiased estimate of $\underline{\theta}$ based on a sample of size *n*.

Definition (2.3.1)

Let x$_1$, x$_2$,…be i.i.d random variables with distribution depending upon a parameter $\underline{\theta}\in\Theta$.

A sequence of estimators $\{\underline{\tilde{\theta}_n}\}$ of $\underline{\theta}$, with $\underline{\tilde{\theta}_n}$ is a function of x$_1$,…, x$_n$, such that $\sqrt{n}(\underline{\tilde{\theta}_n}-\underline{\theta})\xrightarrow{L}N(\underline{0},\mathrm{var}(\underline{\tilde{\theta}}))$.

Whatever be the true value of $\underline{\theta}$, is said to be asymptotically efficient if $\mathrm{var}(\underline{\tilde{\theta}})=B(\underline{\theta})^{-1}$ for all $\underline{\theta}\in\Theta$.

Note that by definition, the MLE is asymptotically efficient under the conditions of theorem (3.2.1).

2.4-Asymptotic Sufficiency of MLE

An explained about the asymptotic theory, one way of explaining the good asymptotic properties of MLE's is via their asymptotic sufficiency. The proof is implicit in the earlier arguments, under the usual regularity conditions of that theorems.

The multiparameter case is immediate from the relationship:

$$\sqrt{n}(\underline{\hat{\theta}}-\underline{\theta})B(\underline{\theta})=\frac{1}{\sqrt{n}}\sum_{1}^{n}\psi(x_i,\underline{\theta})......(*)$$

such that, the relationship (*) between the MLE and the efficient score vector, the asymptotic sufficiency can also be expressed in terms of $\psi(x,\underline{\theta})$.

2.5-Restricted Maximum Likelihood Estimators

In certain problems under one of possible number of assumptions, the vector parameter $\underline{\theta}\in\Theta$ is further restricted to some subspace H of Θ. In that event if we wish to estimate $\underline{\theta}$ under such assumptions we shall need to maximize $f(\underline{x}/\underline{\theta})$ within H instead of Θ.

If $\underline{\hat{\hat{\theta}}}$ is that value that maximizes $f(\underline{x}/\underline{\theta})$ within H

i.e $f(\underline{x},\underline{\hat{\hat{\theta}}}) = \max\limits_{\underline{\theta}\in \mathrm{H}} f(\underline{x}/\underline{\theta})$.

Then $\underline{\hat{\hat{\theta}}}$ is called the restricted MLE (RMLE) of $\underline{\theta}$.

Usually, the restricted space H is described by a set of side conditions, i.e $\underline{\theta}\in\mathrm{H}$ if $\underline{\theta}$ satisfies the side equations:

$$h_1(\underline{\theta})=0, h_2(\underline{\theta})=0,\ldots.,h_r(\underline{\theta})=0, \text{i.e } \underline{h}(\underline{\theta})=\underline{0}$$

One way of finding the RMLE is to use the method of Lagrange multipliers. This method tells us that $\underline{\hat{\hat{\theta}}}$ satisfies:

$$\left.\begin{array}{r} \dfrac{\partial}{\partial\theta_i}\ell_n(\underline{\theta})+\sum\limits_{j=1}^{r}\lambda_j\dfrac{\partial}{\partial\theta_i}h_j(\underline{\theta})=0 \\ h_j(\underline{\theta})=0 \end{array}\right\} \begin{array}{l} i=1,2,..,k \\ j=1,2,..,r \end{array} \quad (r<k)$$

In matrix notation:

$$D_{\underline{\theta}}\ell_n(\underline{\theta})+\mathrm{H}(\underline{\theta})\underline{\lambda}=0$$
$$\underline{h}(\underline{\theta})=0$$

where $D_{\underline{\theta}} = \dfrac{\partial}{\partial \underline{\theta}}$, and

$$\mathrm{H}(\underline{\theta}) = D_{\underline{\theta}}\, \underline{h}^T(\underline{\theta}) = \begin{pmatrix} \dfrac{\partial}{\partial \theta_1} \\ . \\ . \\ . \\ . \\ \dfrac{\partial}{\partial \theta_k} \end{pmatrix} (h_1(\underline{\theta}), \ldots.., h_r(\underline{\theta}))$$

i.e

$[\mathrm{H}(\underline{\theta})]_{i,j} = \dfrac{\partial}{\partial \theta_i} h_j(\underline{\theta})$ and $\underline{\lambda} = \begin{pmatrix} \lambda_1 \\ . \\ . \\ . \\ \lambda_r \end{pmatrix}$ are the Lagrange multipliers.

We assume that rank $(\mathrm{H}(\underline{\theta})) = r$.

<u>Theorem (2.5.1)</u>

Let $\underline{\theta}$ be an identifiable *k*-dimensional parameter of a distribution and let $\hat{\hat{\underline{\theta}}}(\underline{x})$ be the RMLE of $\underline{\theta}$ based on a random sample $\underline{\mathrm{x}}$ from the distribution under the side conditions $\underline{h}(\underline{\theta}) = \underline{0}, \underline{h}$ being *r*-dimensional. Then under the regularity conditions of theorem (3.1.1).

(i) $\hat{\hat{\underline{\theta}}}$ is strongly consistent (i.e $\hat{\hat{\underline{\theta}}} \xrightarrow{a.s} \underline{\theta}$)

(ii) $\hat{\hat{\underline{\theta}}}(\underline{x}) \sim N(\underline{\theta}, \dfrac{p(\underline{\theta})}{n})$ for large sample size *n*, where matrix $p(\underline{\theta})$ satisfies:

$$\begin{pmatrix} B(\underline{\theta}) & \mathrm{H}(\underline{\theta}) \\ \mathrm{H}^T(\underline{\theta}) & 0 \end{pmatrix}^{-1} = \begin{pmatrix} p(\underline{\theta}) & Q(\underline{\theta}) \\ Q^T(\underline{\theta}) & R(\underline{\theta}) \end{pmatrix}$$

i.e $p(\underline{\theta})B(\underline{\theta})p(\underline{\theta}) = p(\underline{\theta})$.

<u>Proof:</u>

(i) For any $\underline{\theta} \in \mathrm{H}$, (i.e $\underline{h}(\underline{\theta}) = \underline{0}$) we have

$Z(\underline{\theta}') < Z(\underline{\theta})$ where $Z = E(\log L_n(\underline{\theta})) = E(\ell_n(\underline{\theta}))$.

If $\underline{\theta}' \neq \underline{\theta}$, and $\frac{1}{n}\ell_n(\underline{\theta}') \xrightarrow{a.s} Z(\underline{\theta}')$, so that $\underline{\hat{\theta}}$, the maximizing value of $\frac{1}{n}\ell_n(\underline{\theta}')$ is close to $\underline{\theta}$, the maximizing point of $Z(\underline{\theta}')$.

Since $\underline{\theta}\in$H, it follows that $\underline{\hat{\theta}}$ is close to H, in which case the maximizing value, $\underline{\hat{\hat{\theta}}}$ of $\frac{1}{n}\ell_n(\underline{\theta}')$ in H, must close to $\underline{\hat{\theta}}$ and hence close to $\underline{\theta}$.

Under the same regularity conditions of theorem (3.1.1) $\underline{\hat{\hat{\theta}}}(x) \xrightarrow{a.s} \underline{\theta}$ as $n \to \infty$ establishing part (i).

(ii) since $\underline{\hat{\hat{\theta}}}$ satisfies

$$\left.\begin{array}{r} D_{\underline{\theta}}\ell_n(\underline{\hat{\hat{\theta}}}) + \mathrm{H}(\underline{\hat{\hat{\theta}}})\,\underline{\lambda} = 0 \\ \underline{h}(\underline{\hat{\hat{\theta}}}) = 0 \end{array}\right\} \qquad \ldots\ldots (1)$$

We have by Taylor's expansion and the fact that $(\underline{\hat{\hat{\theta}}})$ is consistent:

$$\left.\begin{array}{r} D_{\underline{\theta}}\ell_n(\underline{\theta}) + D^2_{\underline{\theta}}\ell_n(\underline{\theta})\,(\underline{\hat{\hat{\theta}}} - \underline{\theta}) + \mathrm{H}(\underline{\theta})\,\underline{\lambda} + D_{\underline{\theta}}(\mathrm{H}(\underline{\theta})\,\underline{\lambda})(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \approx 0 \\ \underline{h}(\underline{\theta}) + D_{\underline{\theta}}\,\underline{h}(\underline{\theta})(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \approx 0 \\ \text{i.e} \quad \mathrm{H}^T(\underline{\theta})\,(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \approx 0 \\ \ldots\ldots\ldots(2) \end{array}\right\}$$

($\approx$ means approximately)

Also from (i), we have [expand the first term around $(\underline{\hat{\theta}})$]

$$D_{\underline{\theta}}\ell_n(\underline{\hat{\theta}}) + D^2_{\underline{\theta}}\ell_n(\underline{\hat{\theta}})\,(\underline{\hat{\hat{\theta}}} - \underline{\hat{\theta}}) + \mathrm{H}(\underline{\hat{\hat{\theta}}})\,\underline{\lambda} \approx 0.$$

That is,

$$D^2_{\underline{\theta}}\ell_n(\underline{\hat{\theta}})\,(\underline{\hat{\hat{\theta}}} - \underline{\hat{\theta}}) + \mathrm{H}(\underline{\hat{\hat{\theta}}})\,\underline{\lambda} \approx 0,$$

thus

$|\mathrm{H}(\underline{\hat{\hat{\theta}}})\underline{\lambda}| = |\underline{\hat{\hat{\theta}}} - \underline{\hat{\theta}}| = 0|\underline{\hat{\hat{\theta}}} - \underline{\theta}|$. So that

$|\mathrm{H}(\underline{\theta})\underline{\lambda}| = 0|\underline{\hat{\hat{\theta}}} - \underline{\theta}|$.

And $|\mathrm{H}(\underline{\theta})\underline{\lambda}(\underline{\hat{\hat{\theta}}} - \underline{\theta})| = 0|\underline{\hat{\hat{\theta}}} - \underline{\theta}|^2$ and therefore can be omitted from (2) to obtain

$$D_{\underline{\theta}}\ell_n(\underline{\theta}) + D^2_{\underline{\theta}}\ell_n(\underline{\theta})\,(\underline{\hat{\hat{\theta}}} - \underline{\theta}) + \mathrm{H}(\underline{\theta})\underline{\lambda} \approx 0$$

$$\mathrm{H}^T(\underline{\theta})\,(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \approx 0$$

or

$$\begin{pmatrix} \frac{-1}{n}D^2_{\underline{\theta}}\ell_n(\underline{\theta}) & \mathrm{H}(\underline{\theta}) \\ \mathrm{H}^T(\underline{\theta}) & 0 \end{pmatrix} \begin{pmatrix} \sqrt{n}(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \\ -\frac{1}{\sqrt{n}}\underline{\lambda} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}}D_{\underline{\theta}}\ell_n(\underline{\theta}) \\ 0 \end{pmatrix}.$$

But

$$\frac{1}{\sqrt{n}}D_{\underline{\theta}}\ell_n(\underline{\theta}) \xrightarrow{L} N(\underline{0}, B(\underline{\theta})), \text{and}$$

$$-\frac{1}{n}D^2_{\underline{\theta}}\ell_n(\underline{\theta}) \xrightarrow{a.s} B(\underline{\theta})$$

Hence for large *n*

$$\begin{pmatrix} \sqrt{n}(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \\ -\frac{1}{\sqrt{n}}\underline{\lambda} \end{pmatrix} \approx \begin{pmatrix} B(\underline{\theta}) & \mathrm{H}(\underline{\theta}) \\ \mathrm{H}^T(\underline{\theta}) & 0 \end{pmatrix}^{-1} \begin{pmatrix} N(\underline{0}, B(\underline{\theta})) \\ 0 \end{pmatrix}$$

or $$\begin{pmatrix} \sqrt{n}(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \\ \frac{-1}{\sqrt{n}}\underline{\lambda} \end{pmatrix} \approx \begin{pmatrix} P(\underline{\theta}) & Q(\underline{\theta}) \\ Q^T(\underline{\theta}) & R(\underline{\theta}) \end{pmatrix} \begin{pmatrix} N(\underline{0}, B(\underline{\theta})) \\ 0 \end{pmatrix},$$

so,

$$\sqrt{n}(\underline{\hat{\hat{\theta}}} - \underline{\theta}) \approx P(\underline{\theta})N(\underline{0}, B(\underline{\theta})),$$

then

$$\underline{\hat{\hat{\theta}}} \sim N(\underline{\theta}, \frac{1}{n}P(\underline{\theta})B(\underline{\theta})P(\underline{\theta})).$$

Since $B(\underline{\theta})$ is symmetric, so is $P(\underline{\theta})$. Also

$$\begin{pmatrix} B(\underline{\theta}) & \mathrm{H}(\underline{\theta}) \\ \mathrm{H}^T(\underline{\theta}) & 0 \end{pmatrix} \begin{pmatrix} P(\underline{\theta}) & Q(\underline{\theta}) \\ Q^T(\underline{\theta}) & R(\underline{\theta}) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \qquad \ldots\ldots\ldots (3)$$

That is

$$B(\underline{\theta})P(\underline{\theta}) + \mathrm{H}(\underline{\theta})Q^T(\underline{\theta}) = I \qquad \ldots\ldots\ldots.(4)$$

$$\mathrm{H}^T(\underline{\theta})P(\underline{\theta}) = 0\,.$$

The latter implies

$$0 = p^T(\underline{\theta})\mathrm{H}(\underline{\theta}) = P(\underline{\theta})\mathrm{H}(\underline{\theta})\,.$$

Now, multiply both sides of (4) by $P(\underline{\theta})$, we get

$$P(\underline{\theta})B(\underline{\theta})P(\underline{\theta}) + P(\underline{\theta})\mathrm{H}(\underline{\theta})Q^T(\underline{\theta}) = P(\underline{\theta})$$

implies

$$P(\underline{\theta})B(\underline{\theta})P(\underline{\theta}) = P(\underline{\theta})$$

and

$\underline{\hat{\hat{\theta}}} \sim N(\underline{\theta}, \frac{P(\underline{\theta})}{n})$ for large *n*, establishing the part (ii).

2.6- Confidence Interval for Large Samples

We know that under very general conditions, the MLE $\hat{\theta}$ is distributed normally about mean θ with variance

$$\mathrm{var}(\hat{\theta}) = \frac{1}{nE[\frac{\partial^2}{\partial\theta^2}\log f(x/\theta)]}.$$

For large samples:

$$p\{\hat{\theta} - Z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\theta})} \le \theta \le \hat{\theta} + Z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\theta})}\} = 1-\alpha$$

where α is the significance level.

REFERENCES

[1] Ash, R. (1972): “Real Analysis and Probability”, Academic Press, New York.

[2] Cox, D.R. and Hinkley, D.V. (1974): “Theoretical Statistics”, Chapman and Hall, London.

[3] Dedwicz, E.J. (1976): “Introduction to Statistics and Probability”, Holt, Rinehart and Winston, U.S.A.

[4] Ferguson, T.S. (1996): “A Course in Large Sample Theory”, Chapman and Hall, Great British. New York.

[5] Klimov, G. (1986): “Probability Theory and Mathematical Statistics” Mir, Publishers. Moscow.

[6] Rao, C.R. (1973): “Linear Statistical Inference and its Applications”. John Wiley and Sons, New York.

[7] Sharma, J.N. (1977): “Topology”, Meerut College, Meerut.

[8] Silvey, S.D. (1978): “Statistical Inference”. Chapman and Hall, London.

[9] Wald A. (1949): “Note on The consistency of the Maximum likelihood Estimate”. Ann.Math.Statist. Vol.20, pp 595-601.

[10] Wolfowitz, J. (1949): “On Walds Proof of the Consistency of the Maximum likelihood Estimate” Ann. Math. Statist.Vol.20, pp. 601-602.