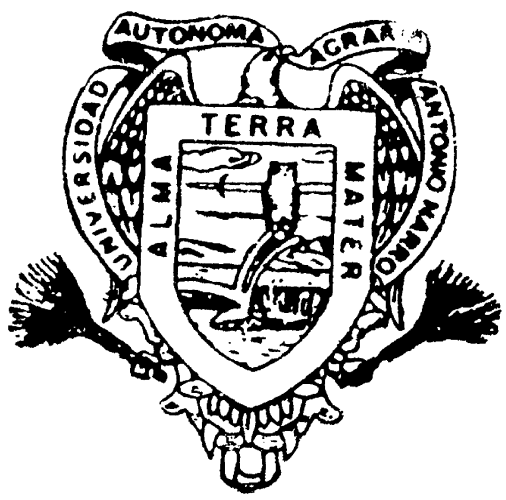
CRITERIO DE CONSISTENCIA PARA ESTIMADORES DE VEROSIMILITUD MAXIMA

EDNA MARINA GONZALEZ MARTINEZ

TESIS

Presentada como Requisito Parcial para
Obtener el Grado de:

MAESTRO EN
ESTADISTICA APLICADA



UNIVERSIDAD AUTONOMA AGRARIA
"ANTONIO NARRO"
PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coahuila, México
Marzo de 2010

Universidad Autónoma Agraria Antonio Narro
Dirección de Postgrado

CRITERIO DE CONSISTENCIA PARA ESTIMADORES DE VEROSIMILITUD MÁXIMA

TESIS

Por:

EDNA MARINA GONZÁLEZ MARTÍNEZ

Elaborada bajo la supervisión del comité particular de asesoría y aprobada como requisito parcial para optar al grado de

MAESTRO EN
ESTADÍSTICA APLICADA

COMITÉ PARTICULAR

Asesor principal: _____
Dr. Rolando Cavazos Cadena

Asesor: _____
Dr. Mario Cantú Sifuentes

Asesor: _____
Dr. Federico Zertuche Luis

_____
Dr. Jerónimo Landeros Flores
Director de Postgrado

Buenavista, Saltillo, Coahuila, Marzo de 2010

AGRADECIMIENTOS

Dr. Mario Cantú Sifuentes

Dr. Rolando Cavazos Cadena

Dr. Federico Zertuche Luis

Dr. Jerónimo Landeros Flores

COMPENDIO

CRITERIO DE CONSISTENCIA PARA ESTIMADORES DE VEROSIMILITUD MÁXIMA

Por

EDNA MARINA GONZÁLEZ MARTÍNEZ

MAESTRÍA EN

ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA

"ANTONIO NARRO"

BUENAVISTA, SALTILLO, COAHUILA, Marzo de 2010

Dr. Rolando Cavazos Cadena –Asesor–

Palabras clave: Ley cero-uno de Hewitt-Savage, Ley fuerte de los grandes números, Criterio de consistencia, Acotamiento casi en toda parte.

Este trabajo considera una sucesión $\{X_i\}$ de objetos aleatorios que son independientes e idénticamente distribuidos. La distribución común de los X_i's es absolutamente continua con respecto a una medida dada y la correspondiente densidad depende de un parámetro desconocido. Bajo condiciones topólogicas y de continuidad poco restrictivas, se obtiene una condición necesaria y suficiente para la consistencia de una sucesión de estimadores de verosimilitud máxima. Al aplicar dicha caracterización al caso en que el espacio de parámetros está contenido en un espacio Euclideano de dimensión finita, la conclusión es que la sucesión es consistente si y sólo si es acotada con probabilidad 1.

ABSTRACT

A CONSISTENCY CRITERION FOR MAXIMUM LIKELIHOOD ESTIMATORS

By

EDNA MARINA GONZÁLEZ MARTÍNEZ

MASTER

APPLIED STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA

ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, March, 2010

Dr. Rolando Cavazos Cadena –Advisor–

Key Words: Hewitt-Savage zero-one law, Strong law of large numbers, Consistency criterion, Boundedness almost everywhere.

A sequence $\{X_i\}$ of independent and identically distributed random objects is considered. The common distribution of the X_i's is absolutely continuous with respect to a given measure, and the corresponding density is not completely specified but depends on an unknown parameter. Under mild topological and continuity requirements, a necessary and sufficient criterion for the consistency of a sequence of maximum likelihood estimators is obtained. When this characterization is applied to the case in which the parameter belongs to a finite dimensional Euclidean space, the conclusion is that *the sequence is consistent if and only if it is bounded* with probability 1.

Índice de Contenido

Casella y Berger, 2001). Por otro lado, es posible construir ejemplos en los que los estimadores de verosimilitud máxima no son consistentes, esto es, a medida que el tamaño de la muestra crece los estimadores no convergen al verdadero valor del parámetro; un ejemplo sencillo ilustrando este comportamiento se presenta en el Capítulo 5, mientras que otros ejemplos más sofisticados pueden encontrarse en Lehmann y Casella (1998) y Bahadur (1958). Por otro lado, un análisis cuidadoso de los argumentos utilizados para demostrar las propiedades de optimalidad asintótica de los estimadores de verosimilitud máxima, pone de manifiesto que un requisito indispensable para obtenerlas es la consistencia de los estimadores (Serfling, 1988), y el hecho de que dicha propiedad falle en algunos casos, proporciona la motivación para el desarrollo de este trabajo, en el cual se analiza el siguiente.

Problema Básico: *Desarrollar un criterio necesario y suficiente para la consistencia de los estimadores de verosimilitud máxima.*

La consistencia de los los estimadores máximo verosímiles ha sido obtenida bajo las siguientes condiciones suficientes impuestas sobre las posibles densidades $f(\mathbf{y}, \theta)$ de Y (Newey y McFadden, 1993):

i) Para cada $\mathbf{y}$, la función $\theta \mapsto \log\left(f(\mathbf{y}, \theta)\right)$ es continua y el espacio de parámetros Θ es compacto;

ii) Para cada $\mathbf{y}$, la función $\theta \mapsto \log\left(f(\mathbf{y}, \theta)\right)$ es estrictamente cóncava, el espacio de parámetros es convexo y el verdadero valor de θ se ubica en el interior de Θ.

En este punto, es importante notar que existen modelos en los que estas condiciones no se satisfacen, y en este caso es importante disponer de un criterio para establecer la consistencia de los estimadores de verosimilitud máxima. El criterio que se obtiene en este trabajo no involucra la estructura específica de las posibles densidades de Y más allá de requisitos estándar de continuidad, sino que se expresa en términos de la sucesión calculada de estimadores. De manera simplificada, la principal contribución de este trabajo, establecida en le Capítulo 5, consiste en establecer el siguiente

Resultado Principal: *La sucesión de estimadores de verosimilitud máxima es consistente si, y sólo si, la sucesión es acotada con probabilidad uno.*

1.2 Los Instrumentos Técnicos Fundamentales

Las herramientas estadísticas que se utilizarán para analizar la consistencia de los estimadores de verosimilitud máxima son, esencialmente, dos:

E1. La ley de los grandes números, y

E2. La desigualdad de Jensen, la cual relaciona las ideas de valor esperado y de función cóncava. Cuando se aplica esta desigualdad a la función logaritmo natural se obtiene la desigualdad de Kullback, que será muy útil en los dos últimos capítulos de este trabajo. Por otro lado, el instrumento matemático esencial en el desarrollo subsecuente es el siguiente:

M1. La propiedad de Heine-Borel de conjuntos compactos en un espacio métrico. (Un conjunto contenido en $\mathbb{R}^k$ es compacto si es cerrado y acotado). Dicha propiedad establece que al cubrir un conjunto compacto mediante una colección de conjuntos abiertos, existe una subcolección finita que también cubre al conjunto (Dugundji 1970, Munkres 1989, Rudin 1968).

1.3 La Organización

La presentación del material subsecuente ha sido organizada de la siguiente manera: En el Capítulo 2 se describe brevemente el problema de estimación puntual y se introducen los modelos paramétricos. Luego, en el Capítulo 3 se introduce formalmente la noción de consistencia que se usará en este trabajo y se estudia el resultado esencial sobre esta idea, a saber, la ley de los grandes números, así como las desigualdades de Jensen y de Kullback. Posteriormente, en el Capítulo 4 se introduce formalmente el método de verosimilitud máxima y se demuestra que, cuando el espacio de parámetros es finito, la consistencia de los estimadores máximo verosimiles puede obtenerse por medio de la desigualdad de Kullback; la idea es ilustrar el papel relevante que dicha desigualdad desempeña en el análisis del esquema de estimación. Finalmente, la exposición concluye en el Capítulo 5 donde se demuestra el principal resultado de este trabajo, a saber, el criterio necesario y suficiente para que una sucesión de estimadores de verosimilitud máxima sea consistente. La presentación en esta parte es autocontenida y relativamente independiente del resto del material. La razón es que el enfoque es bastante general y requiere de un instrumental matemático más avanzado que el empleado en los capítulos precedentes.

CAPÍTULO 2

ESTIMACIÓN PUNTUAL

2.1 MODELOS ESTADÍSTICOS

El material básico de un problema de inferencia estadística es un conjunto de datos $Y = (Y_1, Y_2, \ldots, Y_n)$, y el objetivo del análisis es utilizar esas cantidades para establecer conclusiones objetivas sobre aspectos desconocidos del proceso que genera las observaciones. Para alcanzar ese propósito, el método estadístico inicia suponiendo que los datos observados son variables o vectores aleatorios cuya distribución común no está completamente especificada, pues depende de esos aspectos que no son completamente conocidos para el observador. Si $\mathcal{F}$ denota a la familia de todas las posibles distribuciones de los datos Y, se escribe

$$Y \sim \mathcal{F} \tag{2.1.1}$$

y especificar un modelo estadístico para los datos consiste en establecer cual es la familia $\mathcal{F}$ de posibles distribuciones.

Ejemplo 2.1.1. [Lanzamiento de una moneda.] Considere una moneda que no se sabe si está cargada o no. Se lanza la moneda en n ocasiones, y se observan los resultados $Y_1, Y_2, \ldots, Y_n$ de cada lanzamiento codificados como 1 o 0. En este caso las variables Y_i son independientes y su distribución común satisface

$$P[Y_i = 1] = 1 - P[Y_i = 0] = \theta \in [0, 1]$$

o bien

$$P[Y_i = u] = \theta^u (1-\theta)^{1-u} I_{\{0,1\}}(u),$$

donde θ es desconocido, puesto que no se sabe si la moneda es "legal" o no. La distribución determinada por la anterior relación desplegada es la distribución de Bernoulli con parámetro θ, denotada por $Ber(\theta)$. Debido a que $Y_1, Y_2, \ldots, Y_n$ son independientes con distribución común $Ber(\theta)$, la distribución del vector $Y = (Y_1, Y_2, \ldots, Y_n)$ está dada por

$$\begin{aligned} P[Y = \mathbf{u}] &= \prod_{i=1}^{n} P[Y_i = u_i] \\ &= \prod_{i=1}^{n} \theta^{u_i} (1-\theta)^{1-u_i} I_{\{0,1\}}(u) \\ &= \theta^{\sum_{i=1}^{n} u_i} (1-\theta)^{n-\sum_{i=1}^{n} u_i} \prod_{i=1}^{n} I_{\{0,1\}}(u_i), \end{aligned}$$

la cual es la distribución de Bernoulli de dimensión n, denotada mediante $Ber_n(\theta)$. Por lo tanto, la familia $\mathcal{F}$ de posibles distribuciones del vector de datos $Y = (Y_1, Y_2, \ldots, Y_n)$ es

$$\mathcal{F} = \{ Ber_n(\theta) \mid \theta \in [0,1] \},$$

completando la especificación del modelo estadístico para Y. □

Ejemplo 2.1.2. [Determinación de una longitud.] Considere el problema de determinar una medida, digamos la longitud θ de una cuerda. En este caso es posible que se tenga cierta información *a priori* sobre θ, por ejemplo que la longitud es por lo menos A y que no excede a cierto número B, de manera que

$$\theta \in [A, B].$$

Para encontrar la longitud desconocida, se toman varias mediciones independientes $Y_1, Y_2, \ldots, Y_n$, las cuales, en general, diferirán debido a fluctuaciones aleatorias.

Suponga que los errores de medición $Y_i - \theta, \ldots, Y_n - \theta$ son independientes con una misma distribución, digamos, uniforme en el intervalo $[-1, 1]$, esto es, la densidad de cada error $X = Y_i - \theta$ es

$$f(x) = \frac{1}{2} I_{[-1,1]}(x),$$

y por lo tanto, la densidad de cada observación Y_i es

$$f_{Y_i}(y_i) = \frac{1}{2} I_{[\theta-1,\theta+1]}(y_i).$$

Como las diversas mediciones son independientes con la misma distribución, la densidad del vector $Y = (Y_1, Y_2, \ldots, Y_n)$ es

$$\prod_{i=1}^{n} \frac{1}{2} I_{[\theta-1,\theta+1]}(y_i), \quad \mathbf{y} = (y_1, y_2, \ldots, y_n),$$

la cual es la densidad uniforme en el cubo n-dimensional $[\theta - 1, \theta + 1]^n$, denotada por $\mathcal{U}([\theta - 1, \theta + 1]^n)$, de manera que

$$Y \sim \{\mathcal{U}([\theta - 1, \theta + 1]^n) \mid \theta \in [A, B]\},$$

completando la especificación del modelo estadístico para Y. □

Después de los dos ejemplos anteriores, a continuación se establece la definición formal de un modelo estadístico.

Definición 2.1.1. *Dado un vector aleatorio Y, un modelo estadístico para Y es una familia $\mathcal{F}$ de distribuciones de probabilidad para Y.*

Una vez que se ha especificado un modelo estadístico, el analista procede suponiendo que la verdadera distribución de Y es un miembro de $\mathcal{F}$, y su problema consiste en identificarla. Como es claro a partir de los ejemplos anteriores, los aspectos que se desconocen del proceso que genera los datos aparecen involucrados en la especificación de la familia $\mathcal{F}$.

2.2 Familias de Densidades

La discusión precedente pone de manifiesto que, al formular un modelo estadístico, las posibles distribuciones que conforman la familia $\mathcal{F}$ se especifican naturalmente estableciendo la densidad de cada miembro de $\mathcal{F}$ respecto a una medida de Lebesgue o de conteo. Para clarificar esta observación, considere de nueva cuenta el Ejemplo 2.1.1. En ese contexto, se determinó que la función de probabilidad de Y es

$$f(\mathbf{u},\theta) = \prod_{i=1}^{n} \theta^{\sum_{i=1}^{n} u_i} (1-\theta)^{n-\sum_{i=1}^{n} u_i} \prod_{i=1}^{n} I_{\{0,1\}}(u_i),$$

lo cual significa que la probabilidad de que Y pertenezca a un conjunto $C \subset \mathbb{R}^n$ se evalúa mediante

$$\begin{aligned} P_\theta[Y \in C] &= \sum_{\substack{\mathbf{u}\in C \\ u_i=0 \ o \ u_i=1}} \prod_{i=1}^{n} \theta^{\sum_{i=1}^{n} u_i} (1-\theta)^{n-\sum_{i=1}^{n} u_i} \qquad (2.2.2) \\ &= \sum_{\substack{\mathbf{u}\in C \\ u_i=0 \ o \ u_i=1}} f(\mathbf{u},\theta) \end{aligned}$$

Los posibles valores de Y son todos aquellos vectores $\mathbf{u}$ en $\mathbb{R}^n$ cuyas componentes son cero o uno, y la función $f(\mathbf{u},\theta)$ es la densidad de la distribución de Y respecto a la medida de conteo en ese conjunto bajo el supuesto de que el verdadero valor del parámetro es θ.

Considere ahora el Ejemplo 2.1.2. En este caso se determinó que la distribución de Y cuando la verdadera longitud es θ tiene densidad respecto a la medida de Lebesgue en $\mathbb{R}^n$ dada por

$$f(\mathbf{y},\theta) = \prod_{i=1}^{n} \frac{1}{2} I_{[\theta-1,\theta+1]}(y_i), \quad \mathbf{y} = (y_1, y_2, \ldots, y_n),$$

de manera que la probabilidad de que Y pertenezca a un conjunto $C \subset \mathbb{R}^n$ está dada por

$$P_\theta[Y \in C] = \int_C f(\mathbf{y}, \theta)\, dy. \tag{2.2.3}$$

En (2.2.2) y (2.2.3), $P_\theta[Y \in C]$ es la distribución de Y evaluada en C cuando θ es el valor del parámetro, y las ecuaciones muestran explícitamente que dicha distribución está determinada una vez que se conoce la densidad de la distribución respecto a una medida determinada. Note que en ambos ejemplos el parámetro θ representa lo que se desconoce sobre el proceso que genera los datos, y que en ambos casos θ es un número. Con frecuencia, los aspectos desconocidos que determinan la distribución del vector de datos está representado por un vector de números, como lo muestra el siguiente ejemplo.

Ejemplo 2.2.1. Considere el problema de analizar el efecto de dos diferentes tratamientos sobre una variable de respuesta. Por ejemplo, se trata de determinar la dureza promedio de piezas metálicas después de someterlas a calentamiento a dos temperaturas A y B, donde $A < B$. Denote mediante μ_A y μ_B a las durezas promedio bajo las temperaturas A y B, respectivamente. Para estudiar los valores desconocidos de μ_A y μ_B se someten varias piezas a calentamiento bajo la temperatura A, digamos n_A piezas, y se observan las durezas obtenidas $Y_1, Y_2, \ldots, Y_{n_A}$. Similarmente, n_B piezas se calientan a las temperatura B y se obtienen las durezas $Y_{n_A+1}, Y_{n_A+2}, \ldots, Y_{n_A+n_B}$. De esta forma se tiene que

$$E[Y_i] = \mu_A, \quad i = 1, 2, \ldots, n_A, \quad E[Y_i] = \mu_B, \quad i = n_A + 1, n_A + 2, \ldots, n_A + n_B.$$

Para especificar un modelo estadístico para este problema, note que es posible escribir

$$Y_i = \mu_A + \varepsilon_i, \quad i = 1, 2, \ldots, n_A,$$

$$Y_i = \mu_B + \varepsilon_i, \quad i = n_A + 1, n_A + 2, \ldots, n_A + n_N.$$

donde las variables ε_i son discrepancias o errores aleatorios con media cero. Si suponemos que dichos errores tienen distribución normal con varianza común σ^2 y son independientes, entonces las Y_i son independientes y su densidad es

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i-\mu_A)^2}{2\sigma^2}}, \quad \text{si } i = 1, 2, \ldots, n_A,$$

y

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i-\mu_B)^2}{2\sigma^2}}, \quad \text{si } i = n_A + 1, n_A + 2, \ldots, n_A + n_B,$$

de manera que la densidad del vector de observaciones $Y = (Y_1, Y_2, \ldots, Y_n)$, donde $n = n_A + n_B$, es

$$f(\mathbf{y}, \mu_A, \mu_B, \sigma^2) = \prod_{i=1}^{n_A} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i-\mu_A)^2}{\sigma^2}} \prod_{i=n_A+1}^{n_A+n_B} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i-\mu_B)^2}{\sigma^2}},$$

de donde se desprende que la distribución de Y depende de tres cantidades desconocidas: μ_A y μ_B —que son de interés primordial en el estudio— y σ^2, que representa la variabilidad de las durezas debida a fluctuaciones aleatorias ocasionadas por otros factores distintos a la temperatura que no fueron controlados en el experimento que genera las observaciones. De esta forma, en este ejemplo, la densidad de Y depende de $\theta = (\mu_A, \mu_B, \sigma^2)$, un vector tridimensional. □

Los comentarios precedentes pueden resumirse como sigue: La familia de posibles distribuciones en un modelo estadístico *frecuentemente* se especifica señalando la densidad de las distribuciones respecto a la medida de Lebesgue en $\mathbb{R}^n$, o respecto a la medida de conteo en algún conjunto discreto. Cada una de esas densidades depende de un vector de parámetros θ de cierta dimensión $k \geq 1$, y sus componentes representan aspectos del proceso generador de las observaciones que son importantes para el analista.

Definición 2.2.1. *Un modelo estadístico paramétrico para un vector de datos Y es una familia de densidades*

$$\{f(\mathbf{y}, \theta) \mid \theta \in \Theta\}.$$

respecto a la medida de Lebesgue en $\mathbb{R}^n$, o respecto a la medida de de conteo en un conjunto discreto. El conjunto Θ de posibles valores del vector de parámetros θ se llama espacio de parámetros.

En el Ejemplo 2.1.1, el espacio de parámetros es $\Theta = [0,1]$, en el Ejemplo 2.1.2, $\Theta = [A,B]$, mientras que en el Ejemplo 2.2.1 se tiene que $\theta = (\mu_A, \mu_B, \sigma) \in [0,\infty) \times [0,\infty) \times (0,\infty) = \Theta$. Cuando cada distribución en un modelo estadístico se etiqueta mediante un vector θ de dimensión finita, el modelo se denomina paramétrico. Note que cuando se especifica un modelo paramétrico, la distribución P_θ del vector de datos Y cuando el verdadero valor del parámetro es θ está determinada por

$$P_\theta[Y \in C] = \int_C f(\mathbf{y}, \theta)\, dy \tag{2.2.4}$$

en el caso continuo, mientras que

$$P_\theta[Y \in C] = \sum_{y \in C} f(\mathbf{y}, \theta) \tag{2.2.5}$$

en el caso discreto.

Observación 2.2.1. Es conveniente puntualizar que, además de los modelos paramétricos, existen otro tipo de modelos cuyas distribuciones no se etiquetan naturalmente mediante un vector de dimensión finita. Dichos modelos se denominan *no paramétricos* (Randles y Wolfe, 1989). En este trabajo todos los modelos considerados son paramétricos.

2.3 Estimadores Puntuales

Considere un modelo estadístico paramétrico determinado por la familia de densidades $\{f(\mathbf{y}, \theta) \mid \theta \in \Theta\}$, de manera que la distribución de Y cuando θ es el verdadero valor del parámetro está determinada por (2.2.4) o (2.2.5), así que conocer θ implica conocer completamente la distribución de Y. Recordando que el vector θ incluye los 'aspectos' del proceso que genera los datos que no son conocidos, se sigue que los datos observados Y deben usarse para aproximar al verdadero valor de θ. Con frecuencia, será de interés aproximar no sólo a θ sino a funciones de θ. Este punto se ilustra en los siguientes ejemplos.

Ejemplo 2.3.1. En el Ejemplo 2.1.2, suponga que la cuerda de longitud θ se va a partir en dos segmentos de longitudes $\frac{\theta}{3}$ y $\frac{2\theta}{3}$. Con el segmento más corto se construirá un cuadrado, mientras que con el segmento más largo se formará una circunferencia. El interés en este contexto se centra en el área total encerrada por el cuadrado y la circunferencia. Denotando esta área por $A(\theta)$, esta cantidad se calcula como sigue:

i) Recordando que un cuadrado con perímetro P tiene área $\frac{P^2}{16}$, la figura cuadrada de perímetro $\frac{\theta}{3}$ tiene área $\frac{\theta^2}{144}$;

ii) Aplicando el hecho de que el área de un círculo cuyo perímetro es P está dada por $\pi\left(\frac{P}{2\pi}\right)^2 = \frac{P^2}{4\pi}$, el disco de perímetro $\frac{2\theta}{3}$ tiene área $\frac{\theta^2}{9\pi}$.

Por lo tanto,

$$A(\theta) = \frac{\theta^2}{144} + \frac{\theta^2}{9\pi}$$

es la función de interés cuyo valor se desea estimar mediante los datos. □

Ejemplo 2.3.2. En el Ejemplo 2.2.1, $\theta = (\mu_A, \mu_B, \sigma)$, donde μ_A y μ_B son las durezas de las piezas metálicas al someterlas a las temperaturas A y B, respectivamente. En este caso es interesante comparar las dos durezas y por lo tanto es importante estimar el verdadero valor de

$$D(\theta) = \mu_B - \mu_A.$$

Si las piezas metálicas son los eslabones de una cadena que se unirán, entonces otra cantidad de interés es

$$g(\theta) = \text{mín}\{\mu_A, \mu_B\}.$$

Por supuesto, $g_1(\theta) = \mu_A$ y $g_2(\theta) = \mu_B$ también parecen importantes en este contexto. □

Como estos ejemplos muestran, dado un modelo paramétrico, el interés del investigador se centra en obtener aproximaciones para cierta función $g(\theta)$ del parámetro θ. Las aproximaciones al valor $g(\theta)$ que se obtengan mediante los datos observados $Y_1, \ldots, Y_n$, se llaman estimaciones de $g(\theta)$. Esta idea es formalmente introducida a continuación.

Definición 2.3.1. *Considere un modelo estadístico paramétrico determinado por la familia de densidades* $\{f(\mathbf{y},\theta)\,|\,\theta \in \Theta\}$, *y sea* $g\colon \Theta \to \mathbb{R}$ *una función dada. Un estimador de* $g(\theta)$ *basado en* $Y_1, Y_2, \ldots, Y_n$ *es una función*

$$\hat{g}_n \equiv \hat{g}_n(Y_1, Y_2, \ldots, Y_n)$$

cuyos valores serán utilizados por el observador como una 'aproximación' del valor $g(\theta)$ *asociado al verdadero valor del parámetro. Al observar* $Y = \mathbf{y}$ *se obtiene el valor específico* $g(\mathbf{y})$, *el cual es la estimación correspondiente al evento* $Y = \mathbf{y}$.

Esta idea de estimador es bastante general, y la función $\hat{g}_n$ puede ser producida por cualquier método, pero es razonable requerir que, a medida que el número n de datos observados crezca, las cantidades $\hat{g}_n(\mathbf{Y}_n, \ldots, \mathbf{Y}_n)$ se aproximen al valor $g(\theta_0)$ asociado al verdadero valor del parámetro θ_0.

Ejemplo 2.3.3. En el Ejemplo 2.1.2, $Y_1, \ldots, Y_n$ son diversas mediciones de la longitud desconocida θ de una cuerda. En este contexto, parece razonable estimar θ mediante

$$\hat{\theta} = \overline{Y}_n = \frac{Y_1 + \cdots Y_n}{n}$$

Si se desea estimar la función de área

$$A(\theta) = \frac{\theta^2}{144} + \frac{\theta^2}{9\pi}$$

del Ejemplo 2.3.1, un estimador razonable es

$$\hat{A} = \frac{\hat{\theta}^2}{144} + \frac{\hat{\theta}^2}{9\pi} = \frac{\overline{Y}_n^2}{144} + \frac{\overline{Y}_n^2}{9\pi}$$

Por otro lado, debido a que la distribución de las mediciones Y_i es simétrica respecto a θ, otro estimador razonable de θ es

$$\tilde{\theta} = med(Y_1, Y_2, \ldots, Y_n)$$

y entonces la función de área $A(\theta)$ puede estimarse mediante

$$\tilde{A} = \frac{\tilde{\theta}^2}{144} + \frac{\tilde{\theta}^2}{9\pi} = \frac{med(Y_1, \ldots, Y_n)^2}{144} + \frac{med(Y_1, \ldots, Y_n)^2}{9\pi}$$

Es claro que θ y $A(\theta)$ pueden estimarse de varias maneras. □

Ejemplo 2.3.4. En el Ejemplo 2.2.1 el parámetro es $\theta = (\mu_A, \mu_B, \sigma)$. Como las observaciones $Y_1, \ldots, Yn_A$ tienen distribución simétrica respecto a μ_A, un estimador razonable de μ_A es

$$\hat{\mu}_A = med(Y_1, \ldots, Y_{n_A})$$

y similarmente

$$\hat{\mu}_B = med(Y_{n_A+1}, \ldots, Y_{n_A+n_B}).$$

Entonces, $g(\theta) = \mu_B - \mu_A$ puede estimarse naturalmente mediante

$$\hat{g} = \hat{\mu}_B - \hat{\mu}_A.$$

Es posible construir otros estimadores razonables para las anteriores cantidades. Por ejemplo, como μ_A es la media de las observaciones $Y_1, \ldots, Y_{n_A}$ es natural estimar μ_A mediante

$$\tilde{\mu}_A = \frac{Y_1 + \cdots + Y_{n_A}}{n_A}$$

y similarmente

$$\tilde{\mu}_B = \frac{Y_{n_A+1}, \ldots, Y_{n_A+n_B}}{n_B},$$

lo que conduce a estimar $g(\theta) = \mu_B - \mu_A$ mediante

$$\tilde{g} = \tilde{\mu}_B - \tilde{\mu}_A = \frac{Y_{n_A+1}, \ldots, Y_{n_A+n_B}}{n_B} - \frac{Y_1 + \cdots + Y_{n_A}}{n_A}$$

□

2.4 CONSISTENCIA

En general, es posible construir diversos estimadores para una función de interés. Más aún, para cada entero n, el estimador $\hat{g}_n = \hat{g}_n(Y_1, \ldots, Y_n)$ será, en general, diferente. El método que genera a los estimadores $\hat{g}_n$, o la sucesión misma de estimadores $\{\hat{g}_n\}$, se denomina consistente si, conforme el número de datos observados n se incrementa, los valores de $\hat{g}_n$ convergen, en algún sentido, al valor desconocido $g(\theta)$ (Casella y Berger, 2001, Mood *et. al.* 1987, Dudewicz y Mishra 1989). Esta propiedad de consistencia es, sin duda, la más básica de las que se pueden requerir de un método de estimación. En la práctica, significa que el esfuerzo temporal o económico que se realiza para generar más y más datos, se ve recompensado con aproximaciones cada vez más cercanas al valor desconocido $g(\theta)$; en ausencia de esta propiedad, sería imposible convencer a alguien de la conveniencia de invertir recursos para aumentar el número de datos observados, y puede decirse que el requisito mínimo para que un método de estimación sea 'razonable' es que sea consistente. *El tema central de este trabajo es la noción de consistencia,* la cual será definida de manera precisa en el siguiente capítulo.

Sin duda alguna, el más fundamental de los resultados de consistencia es la ley de los grandes números, la cual establece que al observar variables aleatorias independientes $Y_1, Y_2, \ldots, Y_n$ con distribución común y esperanza finita μ, entonces $\hat{\mu}_n = \frac{(Y_1 + \ldots Y_n)}{n}$ —la media muestral basada en los n datos— converge hacia μ a medida que n aumenta. Hay varias versiones de este resultado, conocidas como la ley débil y la ley fuerte de los grandes números. La ley débil establece que, para cada $\varepsilon > 0$, conforme n crece la probabilidad de observar que $|\hat{\mu}_n - \mu| > \varepsilon$ se aproxima a cero, mientras que la ley fuerte dice que al incrementarse n, $\hat{\mu}_n$ converge a μ con probabilidad uno (Ash 1972, Dudley 2002). Como se verá posteriormente, este resultado fundamental, que se encuentra ligado a la propia interpretación frecuencial de probabilidad, es el instrumento básico para estudiar la consistencia de estimadores de verosimilitud máxima.

CAPÍTULO 3

LA NOCIÓN DE CONSISTENCIA

3.1 INTRODUCCIÓN

Este capítulo trata sobre la idea central en este trabajo, a saber, el concepto de consistencia de una sucesión de estimadores puntuales. Como punto de partida, sea $Y_1, Y_2, \ldots$ una sucesión de vectores aleatorios independientes con una distribución común, la cual se supone que tiene densidad $\rho(\mathbf{y};\theta)$ respecto a una medida fija. En la aplicaciones, dicha medida es la de Lebesgue, en el caso continuo, o una medida de conteo, en el caso discreto. Por otro lado, θ es un parámetro desconocido, el cual pertenece a un conjunto Θ contenido en R^k para algún entero fijo $k \geq 1$. De esta manera, el observador no conoce exactamente la densidad de la distribución de los vectores Y_i, pero si sabe que pertenece a la familia $\{f(\mathbf{y};\theta) \mid \theta \in \Theta\}$. En el desarrollo subsecuente, $\theta_0 \in \Theta$ *denota al verdadero valor del parámetro*, de manera que la distribución común de los vectores Y_i tiene densidad $f(\mathbf{y};\theta_0)$; sin embargo, θ_0 no es conocido por el observador, y su objetivo es utilizar los datos observados para estimar el valor de θ_0 o, más generalmente, de una función $g(\theta_0)$. Como se mencionó en el capítulo precedente, un estimador de $g(\theta_0)$ basado en $Y_1, Y_2, \ldots, Y_n$ es una función

$$\hat{g}_n \equiv \hat{g}_n(Y_1, Y_2, \ldots, Y_n)$$

la cual será utilizada por el observador como una 'aproximación' de $g(\theta_0)$. Esta idea de estimador es bastante general, y la función $\hat{g}_n$ puede ser producida por cualquier método, pero es razonable requerir que, a medida que el número n de

datos observados crece, los valores de $\hat{g}_n(\mathbf{Y}_n, \ldots, \mathbf{Y}_n)$ se aproximen a $g(\theta_0)$, en cuyo caso la sucesión de estimadores $\{\hat{g}_n\}$ se denomina *consistente*. Esta idea puede formalizarse de varias maneras, y en este trabajo se adoptará la siguiente.

Definición 3.1.1. *Sea $Y_1, Y_2, Y_3 \ldots$ una sucesión de vectores aleatorios independientes e idénticamente distribuidos (i.i. d.), y suponga que la distribución común de las Y_i's tiene densidad $\rho(\mathbf{y}; \theta)$ donde $\theta \in \Theta$. Denote mediante P_θ la distribución correspondiente a $\rho(\mathbf{y}; \theta)$. Dada una función $g(\theta)$, una sucesión $\{\hat{g}_n(Y_1, \ldots, Y_n)\}$ de estimadores de $g(\theta)$ es consistente si, para cada $\theta \in \Theta$,*

$$P_\theta \left[\lim_{n \to \infty} \hat{g}_n(Y_1, \ldots, Y_n) = g(\theta) \right] = 1.$$

Esta noción se denomina consistencia fuerte en la literatura, para distinguirla de otra noción relacionada, llamada consistencia débil o consistencia en probabilidad, la cual require que, para cada $\varepsilon > 0$ y $\theta \in \Theta$,

$$\lim_{n \to \infty} P_\theta \left[\left| \hat{g}_n(Y_1, \ldots, Y_n) - g(\theta) \right| > \varepsilon \right] = 0.$$

Para detalles sobre estas nociones vea, por ejemplo, Dudewicz y Mishra (1989), Mood *et. al.* (1987), o Lehmann y Casella (1998). Puede demostrarse que si una sucesión $\{\hat{g}_n\}$ es consistente en el sentido de la Definición 3.1.1, entonces es consistente en probabilidad, de manera que la noción en la definición anterior es, efectivamente, más fuerte que la idea de consistencia en probabilidad, o débil. En este capítulo se estudia el resultado más básico sobre consistencia, a saber, la ley de los grandes números, y posteriormente se analizan las desigualdades de Jensen y Kullback, herramientas que desempeñan un papel central en el estudio del método de verosimilitud máxima en el siguiente capítulo. La exposición ha sido organizada de la siguiente manera: En la Sección 3.2 se presenta la ley de los grandes números, un importante resultado clásico que establece que los promedios muestrales de variables *i.i. d.* forman una sucesión consistente de estimadores de la media poblacional. En la Sección 3.3 se estudia la desigualdad de Jensen, la cual relaciona el valor esperado de variables aleatorias con la noción de concavidad de una función. El objetivo de esa sección es establecer que, para una variable aleatoria positiva X, la desigualdad $\log E[X] > E\left[\log(X)\right]$ se satisface cuando X no es constante con probabilidad 1, relación que conduce a la desigualdad de Kullback.

3.2 Ley de los Grandes Números

La ley de los grandes números, enunciada a continuación, es el teorema más básico sobre consistencia. Dicho resultado establece que la sucesión de medias muestrales estima consistentemente a la media poblacional.

Teorema 3.2.1. *Sea* $(\Omega, \mathcal{F}, P)$ *un espacio de probabilidad y considere una sucesión de variables aleatorias* $\{X_i : \Omega \to \mathbb{R}\}$ *con las siguientes propiedades:*

i) $X_1, X_2, X_3, \ldots$ *son independientes;*

ii) $X_1, X_2, X_3, \ldots$ *tienen una distribución común, y*

iii) El valor esperado de X_i *es finito, digamos* $\mu = E[X_i]$.

iv) En este caso, existe un evento $\Omega^* \subset \Omega$ *tal que*

$$P[\Omega^*] = 1 \quad y \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} = \mu, \quad \omega \in \Omega^*. \tag{3.2.1}$$

Una demostración de este resultado puede encontrarse, por ejemplo, en Ash (1972) o en Casella y Beger (2001). Usualmente, la conclusión (3.2.1) se expresa escribiendo

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \mu, \quad \text{c.s.}$$

donde c.s. significa 'casi seguramente', otra forma de decir que la propiedad indicada ocurre sobre un evento de probabilidad 1. En el trabajo subsecuente se usará una versión ligeramente más general de la ley de los grandes números. Para establecerla, es conveniente profundizar en uno de los supuestos del Teorema 3.2.1, a saber, la condición de que las variables aleatorias X_i tengan esperanza finita. Considere una variable aleatoria X y defina

$$X^+ = \text{máx}\{X, 0\}, \quad \text{y} \quad X^- = \text{máx}\{-X, 0\}; \tag{3.2.2}$$

de modo que

$$X = X^+ - X^-; \tag{3.2.3}$$

X^+ y X^- son la parte positiva y negativa de X, respectivamente. Si X tiene densidad $f(x)$, entonces

$$E[X^+] = \int_0^\infty x f(x)\, dx, \quad \text{y} \quad E[X^-] = -\int_{-\infty}^0 x f(x)\, dx,$$

mientras que si X es discreta fórmula similares se aplican con las integrales sustituidas por sumatorias. En general, sin importar la naturaleza de la variable aleatoria X,

$$E[X^+] = \int_0^\infty (1 - F(x))\, dx, \quad \text{y} \quad E[X^-] = \int_{-\infty}^0 F(x)\, dx,$$

donde $F(x)$ es la función de distribución de X. *La esperanza de X está definida* cuando

$$E[X^+] < \infty \quad \text{o} \quad E[X^-] < \infty \tag{3.2.4}$$

y en este caso, por definición,

$$E[X] = E[X^+] - E[X^-]; \tag{3.2.5}$$

vea, por ejemplo Ash (1972), o Dudley (2002). Note que la condición (3.2.4) es necesaria para evitar que la fórmula para $E[X]$ en (3.2.5) resulte en '$\infty - \infty$', expresión que carece de sentido. Cuando $E[X^+] = \infty$ y $E[X^-] < \infty$, de acuerdo a la convención usual de que

$$\infty - a = \infty \quad \text{para todo } a \in \mathbb{R},$$

la fórmula (3.2.5) arroja que $E[X] = \infty$. La única forma en que $E[X]$ sea finita, es que

$$E[X^+] < \infty \quad \text{y} \quad E[X^-] < \infty, \tag{3.2.6}$$

pues en estas condiciones $E[X]$ en (3.2.5) es la diferencia de dos números reales (i.e., finitos). El supuesto de que las variables aleatorias X_i en el Teorema 3.2.1 tiene esperanza finita puede escribirse entonces como $E[X_i^+] < \infty$ y $E[X_i^-] < \infty$. Para los propósitos de este trabajo, es conveniente establecer la siguiente forma de la ley de los grandes números, en la cual el supuesto de que la esperanza de las variables X_i sea finito se relaja, imponiendo sólo la condición de que la esperanza de las variables X_i esté definida.

Teorema 3.2.2. *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y considere una sucesión de variables aleatorias $\{X_i : \Omega \to \mathbb{R}\}$ con las siguientes propiedades:*

i) $X_1, X_2, X_3, \ldots$ son independientes;

ii) $X_1, X_2, X_3, \ldots$ tienen una distribución común, y

iii) El valor esperado de X_i está definido, digamos $\mu = E[X_i]$.

En este caso, existe un evento $\Omega^* \subset \Omega$ tal que

$$P[\Omega^*] = 1 \quad y \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} = \mu, \quad \omega \in \Omega^*. \tag{3.2.7}$$

Note que la única diferencia entre las condiciones de los Teorema 3.2.1 y 3.2.2 es que en este último teorema la esperanza común de las variables X_i puede ser ∞ o $-\infty$. Si, por ejemplo, $E[X_i] = \infty$, entonces el Teorema 3.2.2 asegura que $\frac{\lim_{n\to\infty}(X_1+\cdots X_j n)}{n} = \infty$ c.p. 1. Como las variables X_i tienen distribución común, se desprende que $E[X_i^+]$ es la misma para todo i y, similarmente, $E[X_i^-]$ no depende de i.

Demostración del Teorema 3.2.2. Primeramente, observe que es suficiente demostrar el teorema en el caso en que $\mu = E[X_i] = -\infty$ o $\mu = E[X_i] = \infty$, pues el caso en que $E[X_i]$ es finita ya está cubierto por el Teorema 3.2.1. Con esto en mente, suponga que $E[X_i] = -\infty$, esto es, que

$$\mu^+ = E[X_i^+] < \infty \quad y \quad E[X_i^-] = \infty. \tag{3.2.8}$$

En este caso, para cada $N = 1, 2, 3, \ldots$, defina $\tilde{X}_{i,N}$ mediante

$$\tilde{X}_{i,N} := \text{mín}\{X_i^-, N\}. \tag{3.2.9}$$

como $X_i^- \geq 0$ (vea (3.2.2)) se tiene que

$$0 \leq \tilde{X}_{i,N} \leq N,$$

y entonces

$$\tilde{\mu}_N := E[\tilde{X}_{i,N}] \leq N. \tag{3.2.10}$$

Por otro lado, a partir de la especificación de $\tilde{X}_{i,N}$, se desprende que

$$\tilde{X}_{i,N} \leq \tilde{X}_{i,N+1} \leq X_i^-, \quad N = 1, 2, 3, \ldots \quad \text{y} \quad \lim_{N\to\infty} \tilde{X}_{i,N} = X_i^- \tag{3.2.11}$$

y a partir del teorema de convergencia monótona se concluye que

$$\lim_{N\to\infty} \tilde{\mu}_N = \lim_{N\to\infty} E[\tilde{X}_{i,N}] = E[X_i^-] = \infty. \tag{3.2.12}$$

Defina ahora

$$X_{i,N} = X_i^+ - \tilde{X}_{i,N}. \tag{3.2.13}$$

Para cada N fijo, las variables aleatorias $X_{i,N}$ son independientes con distribución común, y $E[X_{i,N}] = E[X_i^+ - \tilde{X}_{i,N}] = \mu^+ - \tilde{\mu}_N$ es finita; vea (3.2.8) y (3.2.10). Luego, aplicando el Teorema 3.2.1, se desprende que existe un evento Ω_N^* con

$$P[\Omega_N^*] = 1 \tag{3.2.14}$$

tal que

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_{i,N}(\omega)}{n} = \mu^+ - \tilde{\mu}_N, \quad \omega \in \Omega_N^*. \tag{3.2.15}$$

Por otro lado, usando que $X_i^- \geq \tilde{X}_{i,N}$ (vea (3.2.11)), se desprende que

$$X_{i,N} = X_i^+ - \tilde{X}_{i,N} \geq X_i^+ - X_i^- = X_i$$

donde la primera y segunda igualdades se deben a (3.2.13) y (3.2.3), respectivamente. Por lo tanto, la desigualdad

$$\frac{\sum_{i=1}^{n} X_i(\omega)}{n} \leq \frac{\sum_{i=1}^{n} X_{i,N}(\omega)}{n}$$

es siempre válida, de tal manera que

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} \leq \limsup_{n} \frac{\sum_{i=1}^{n} X_{i,N}(\omega)}{n}$$

Usando (3.2.15) esta relación conduce a

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} \leq \mu^+ - \tilde{\mu}_N, \quad \omega \in \Omega_N^*. \tag{3.2.16}$$

Para concluir, defina $\Omega^* = \bigcap_{N=1}^{\infty} \Omega_N^*$. En este caso, (3.2.14) implica que

$$P[\Omega^*] = 1$$

mientras que combinando la inclusión $\Omega^* \subset \Omega_N^*$ con (3.2.16) se desprende que

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} \leq \mu^+ - \tilde{\mu}_N, \quad \omega \in \Omega^*, \quad N = 1,2,3,\ldots$$

Tomando límite conforme N tiende a infinito en el lado derecho de esta desigualdad se obtiene que

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} \leq \lim_{N\to\infty} [\mu^+ - \tilde{\mu}_N] = \mu^+ - \infty = -\infty, \quad \omega \in \Omega^*,$$

lo cual equivale a

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} = -\infty = \mu, \quad \omega \in \Omega^*.$$

Como $P[\Omega^*] = 1$, esto muestra que con probabilidad 1 la media muestral converge a la media poblacional conforme n tiende a infinito. El caso $\mu = \infty$ se analiza de forma similar. □

3.3 Desigualdad de Jensen

En esta sección se establece otro de los instrumentos que desempeñan un papel central en el estudio de la consistencia de los estimadores de verosimilitud máxima. Como punto de partida, recuerde que una función $g(x)$ es *cóncava* en un intervalo I si para cada par de puntos x_1 y x_2 en I y para cada $t \in [0,1]$,

$$g(tx_1 + (1-t)x_2) \geq tg(x_1) + (1-t)g(x_2). \tag{3.3.17}$$

Si g es una función derivable, un criterio simple para verificar la concavidad de g es el siguiente: Si $g''(x) \leq 0$ en cada punto de I, entonces g es cóncava en I (Fulks, 1981, Rudin 1968). Si $x_1, x_2, \ldots, x_n$ son puntos en I, un argumento de inducción iniciando con (3.3.17) permite demostrar que

$$g(t_1x_1 + t_2x_2 + \cdots t_nx_n) \geq t_1g(x_1) + t_2g(x_2) + \cdots + t_ng(x_n)$$

siempre que $t_1, t_2, \ldots, t_n$ sean números no negativos cuya suma es la unidad. Considere ahora una variable aleatoria X tal que $P[X = x_i] = t_i$, $i = 1,2,\ldots,n$. En este caso $E[X] = t_1x_1 + t_2x_2 + \cdots t_nx_n$ y $E[g(X)] = t_1g(x_1) + t_2g(x_2) + \cdots t_ng(x_n)$, de manera que la anterior desigualdad desplegada equivale a $g(E[X]) \geq E[g(X)]$, la cual es la *desigualdad de Jensen*. Esta relación se generaliza en el siguiente teorema, en el cual la naturaleza de X es totalmente arbitraria, y sólo se supone que sus valores pertenecen al intervalo I en el cual la función g es cóncava.

Teorema 3.3.1. *Considere un intervalo I de $\mathbb{R}$ y sea $g\colon I \to \mathbb{R}$ una función cóncava en I. Suponga que X es una variable aleatoria tal que $P[X \in I] = 1$, y que $E[X]$ es finita. En este caso, el valor esperado de $g(X)$ existe y satisface*

$$g(E[X]) \geq E\big[g(X)\big].$$

En particular, si $P[X > 0] = 1$, entonces

$$\log\left(E[X]\right) \geq E\big[\log(X)\big].$$

Para una demostración de este resultado vea, por ejemplo, Ash (1972) o Dudewicz y Mishra (1989). En el análisis de la consistencia de los estimadores de verosimilitud máxima se utilizará una caso especial del Teorema 3.3.1, el cual involucra la idea de función *estrictamente cóncava*: Una función definida en un intervalo I es estrictamente cóncava si para cada $x_1, x_2 \in I$ con $x_1 \neq x_2$ y para cada $t \in (0,1)$, la desigualdad estricta ocurre en (3.3.17), esto es, $g\big(tx_1 + (1-t)x_2\big) > tg(x_1) + (1-t)g(x_2)$. Si la función g tiene segunda derivada en I, entonces g es estrictamente cóncava en I si $g''(x) < 0$ en cada punto $x \in I$. En particular, la función $g(x) = \log(x)$ es estrictamente cóncava en $I = (0, \infty)$, pues $g''(x) = \frac{-1}{x^2} < 0$ para todo $x > 0$.

Teorema 3.3.2. *Suponga que g es una función estrictamente cóncava en un intervalo I. Sea X en variable aleatoria para la cual $P[X \in I] = 1$ y cuya esperanza μ es finita. Si X no es constante con probabilidad 1, esto es, si $P[X = \mu] < 1$, entonces*

$$g(E[X]) > E\big[g(X)\big].$$

Una demostración de este resultado puede verse, por ejemplo, en Rudin (1968), Dudley (2001 o en Ash(1972). El siguiente caso especial del Teorema 3.3.2, el cual se obtiene tomando $g(x) = \log(x)$ para $x \in (0, \infty) = I$, desempeñará un papel central en el estudio del método de verosimilitud máxima.

Corolario 3.3.1. *Sea X tal que $P[X \geq 0] = 1$ y $E[X] = \mu \leq 1$. En este caso, si $P[X = 1] < 1$, entonces*

$$0 > E\big[\log(X)\big].$$

Este resultado será utilizado cuando X es el cociente de dos densidades, como se muestra en el siguiente ejemplo.

Ejemplo 3.3.1. [Cociente de densidades.] Sea $\mathbf{Y}$ un vector aleatorio y sean densidad $f_0(\mathbf{y})$ y $f_i(\mathbf{y})$ dos posibles densidades de $\mathbf{Y}$, las cuales se supone que inducen distribuciones distintas, esto es, $\int_A f_0(\mathbf{y}), d\mathbf{y} \neq \int_A f_1(y)\, d\mathbf{y}$ para alguna región A. Denote mediante P_i a la distribución asociada a f_i, $i = 0, 1$, y mediante $E_i[\cdot]$ al operador de valor esperado asociado con P_i, $i = 0, 1$. El objetivo de este ejemplo es verificar que

$$E_0\left[\log\left(\frac{f_1(\mathbf{Y})}{f_0(\mathbf{Y})}\right)\right] = \int \log\left(\frac{f_1(\mathbf{y})}{f_0(\mathbf{y})}\right) f_0(\mathbf{y})\, d\mathbf{y} < 0, \tag{3.3.18}$$

relación conocida como desigualdad de Kullback. Con este fin, sean $\mathcal{Y}_0$ y $\mathcal{Y}_1$ los soportes de f_0 y f_1, respectivamente, esto es,

$$\mathcal{Y}_0 = \{\mathbf{y}\colon f_0(\mathbf{y}) > 0\}, \qquad \mathcal{Y}_1 = \{\mathbf{y}\colon f_1(\mathbf{y}) > 0\},$$

y note que

$$\begin{aligned} E_0\left[\frac{f_1(\mathbf{Y})}{f_0(\mathbf{Y})}\right] &= \int_{\mathcal{Y}_0} \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} f_0(\mathbf{y})\, d\mathbf{y} \\ &= \int_{\mathcal{Y}_0 \cap \mathcal{Y}_1} \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} f_0(\mathbf{y})\, d\mathbf{y} \\ &= \int_{\mathcal{Y}_0 \cap \mathcal{Y}_1} f_1(\mathbf{y})\, d\mathbf{y} \leq \int_{\mathcal{Y}_1} f_1(\mathbf{y})\, d\mathbf{y} = 1, \end{aligned}$$

donde la segunda igualdad se debe al hecho de que $f_1(\mathbf{y})$ se anula para $\mathbf{y} \notin \mathcal{Y}_1$. Esto muestra que

$$X = \frac{f_1(Y)}{f_0(Y)} \tag{3.3.19}$$

satisface $E_0[X] = \mu \leq 1$. Ahora se verificará que $P_0[X = 1] < 1$. El argumento es por contradicción. En efecto, si $P_0[X = 1] = 1$, entonces $X = \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} = 1$ en

el soporte $\mathcal{Y}_0$ de $f_0(Y)$, esto es, $f_1(\mathbf{y}) = f_0(\mathbf{y})$ en $\mathcal{Y}_0$, y entonces $\int_{\mathcal{Y}_0} f_1(\mathbf{y})\,d\mathbf{y} = \int_{\mathcal{Y}_0} f_0(\mathbf{y})\,d\mathbf{y} = 1$, y por lo tanto, $\int_{\mathcal{Y}_0^c} f_1(\mathbf{y})\,d\mathbf{y} = 0$, lo cual significa que, para toda región A,

$$\int_A f_1(\mathbf{y})\,d\mathbf{y} = \int_{A\cap\mathcal{Y}_0} f_1(\mathbf{y})\,d\mathbf{y} = \int_{A\cap\mathcal{Y}_0} f_0(\mathbf{y})\,d\mathbf{y} = \int_A f_0(\mathbf{y})\,d\mathbf{y},$$

y entonces $f_0 = (\mathbf{y})$ y $f_1(\mathbf{y})$ inducen la misma distribución, lo cual contradice la hipótesis de que f_0 y f_1 determinan distribuciones diferentes. Por lo tanto,

$$E_0\left[\frac{f_1(Y)}{f_0(Y)}\right] \leq 1, \quad \text{y} \quad P_0\left[\frac{f_1(Y)}{f_0(Y)} = 1\right] < 1$$

de manera que el corolario precedente implica que (3.3.18) es válida. □

En el siguiente ejemplo se proporciona un cálculo concreto acerca del logaritmo de un cociente de densidades.

Ejemplo 3.3.2. Ahora se proporcionará un cálculo concreto referente a la conclusión del ejemplo previo. Considere una variable aleatoria Y con densidad $f_0(y) = \frac{1}{2}e^{-|y|}, \quad y \in \mathbb{R}$, de manera que

$$E_0[H(Y)] = \int_{\mathbb{R}} H(y) f_0(y)\,dy.$$

Ahora sea $f_1(\mathbf{y})$ la densidad

$$f_1(y) = \frac{1}{2}e^{-|y-\theta|}, \quad y \in \mathbb{R},$$

donde $\theta \in \mathbb{R}$ es arbitrario. En estas circunstancia

$$\log\left(\frac{f_1(Y)}{f_0(Y)}\right) = \log\left(e^{-|Y-\theta|+|Y|}\right) = -|Y-\theta| + |Y|$$

y lo que la conclusión del ejemplo anterior establece en este caso es que

$$0 > \int_{\mathbb{R}} (-|y-\theta| + |y|)\frac{1}{2}e^{-|y|}\,dy.$$

□

Capítulo 4

Método de Verosimilitud Máxima

4.1 Introducción

En este capítulo se introduce el método de verosimilitud máxima para construir estimadores. Sean $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ vectores aleatorios independientes con densidad común $f(\mathbf{y}, \theta)$, donde $\theta \in \Theta$, de manera que la densidad de $\mathbf{Y}_n = (Y_1, \ldots Y_n)$ está dada por

$$f_{\mathbf{Y}_n}(\mathbf{y}_1, \ldots, \mathbf{y}_n, \theta) = \prod_{i=1}^{n} f(\mathbf{y}_i, \theta). \tag{4.1.1}$$

El observador sabe que el verdadero valor del parámetro, digamos θ_0 es un miembro de Θ, pero desconoce el valor exacto de θ_0. Una vez que los datos $Y_1 = \mathbf{y}_1, \ldots, Y_n = \mathbf{y}_n$ han sido observados, la tarea del analista consiste en usar esas observaciones para construir una estimación para el valor desconocido θ_0. A continuación se introduce la idea de función de verosimilitud.

Definición 4.1.1. *Dadas las observaciones* $\mathbf{y}_1, \ldots, \mathbf{y}_n$, *la función de verosimilitud basada en los n datos está dada por*

$$V_n(\theta, \mathbf{y}_1, \ldots, \mathbf{y}_n) = \prod_{i=1}^{n} f(\mathbf{y}_i, \theta), \quad \theta \in \Theta. \tag{4.1.2}$$

Cuando los datos son discretos, $V_n(\theta, \mathbf{y}_1, \ldots, \mathbf{y}_n)$ es la probabilidad de volver a observar $\mathbf{y}_1, \ldots, \mathbf{y}_n$ bajo el supuesto de que θ es el verdadero valor del parámetro. Con esto en mente, es razonable estimar el verdadero valor del parámetro mediante el valor que asigna la mayor probabilidad a los datos observados.

Definición 4.1.2. *Dadas las observaciones $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, el estimador de máxima verosimilitud de θ es el valor $\hat{\theta}_n = \hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in \Theta$ que satisface*

$$V_n(\hat{\theta}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \geq V_n(\theta, \mathbf{Y}_1, \ldots, \mathbf{Y}_n), \quad \theta \in \Theta.$$

Note que $\hat{\theta}_n$ existe cuando Θ es finito, pero en otro caso, la existencia de un maximizador de la función de verosimilitud requiere condiciones adicionales como, por ejemplo, la continuidad de $V_n(\theta, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ respecto a θ y compacidad del espacio de parámetros. Para analizar la consistencia de los estimadores de verosimilitud máxima se supondrá que la siguiente condición se satisface.

Hipótesis 4.1.1. [Identificabilidad.] *Densidades correspondientes a parámetros distintos son diferentes, i.e., si $\theta \neq \theta_1$, entonces existe un conjunto A tal que*

$$\int_A f(\mathbf{y}; \theta)\, d\mathbf{y} \neq \int_A f(\mathbf{y}; \theta_1)\, d\mathbf{y}.$$

En el siguiente capítulo se establecerá un criterio necesario y suficiente para garantizar la consistencia del método de verosimilitud máxima. Por ahora, se establecerá su consistencia bajo el supuesto de que el espacio de parámetros es finito. La idea es apreciar la importancia de la desigualdad de Kullback en los argumentos, sin que su papel quede obscurecido por los detalles técnicos del caso general. Así, el principal objetivo del desarrollo subsecuente es establecer el siguiente resultado.

Teorema 4.1.1. *Suponga que el espacio de parámetros Θ es finito, y que la anterior Hipótesis 4.1.1 es válida. En este contexto, la sucesión $\{\hat{\theta}_n\}$ de estimadores de verosimilitud máxima estima consistentemente a θ, esto es,*

$$P_\theta\left[\lim_{n\to\infty} \hat{\theta}_n = \theta\right] = 1.$$

Este resultado se demostrará después de los preliminares técnicos de la siguiente sección.

4.2 Resultados Auxiliares

El objetivo de esta sección es establecer el siguiente resultado auxiliar, que será utilizado posteriormente para demostrar el Teorema 4.1.1.

Teorema 4.2.1. *Sea $\theta_0 \in \Theta$ un parámetro arbitrario pero fijo, y suponga que la Hipótesis 4.1.1 es válida. En este caso, las siguientes afirmaciones i)– iii) son válidas.*

i) Para cada $\theta \neq \theta_0$,

$$E_{\theta_0}\left[\log\left(\frac{f(\mathbf{Y};\theta)}{f(\mathbf{Y};\theta)}\right)\right] =: \nu(\theta) < 0.$$

ii) Con probabilidad 1 respecto a P_{θ_0},

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}{V_n(\theta_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}\right) = \nu(\theta)$$

iii) Dado $\theta \in \Theta$, existe un evento Ω^* tal que $P_{\theta_0}[\Omega^*] = 1$ para el cual la siguiente afirmación es cierta: Para cada trayectoria $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \ldots)$ en Ω^* existe un entero $N = N(Y,\theta)$ tal que

$$\log V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n) < \log V_n(\theta_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_n), \quad n > N(Y,\theta)$$

o, equivalentemente,

$$V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n) < V_n(\theta_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_n), \quad n > N(Y,\theta).$$

Demostración del Teorema 4.2.1 .

i) Esta parte se desprende del Ejemplo 3.3.1 con $f(\mathbf{y};\theta)$ y $f(\mathbf{y};\theta_0)$ en vez de $f_1(\mathbf{y})$ y $f_0(\mathbf{y})$, respectivamente.

ii) Note que

$$\log\left(\frac{V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}{V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}\right) = \log\left(\frac{\prod_{i=1}^{n} f(\mathbf{Y}_i;\theta)}{\prod_{i=1}^{n} f(\mathbf{Y}_i;\theta_0)}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{f(\mathbf{Y}_i;\theta)}{f(\mathbf{Y}_i;\theta_0)}\right). \qquad (4.2.3)$$

Por otro lado, las variables aleatorias $X_i = \log\left(\frac{f(\mathbf{Y}_i;\theta)}{f(\mathbf{Y}_i;\theta_0)}\right)$ son independientes e idénticamente distribuidas con respecto a P_{θ_0}, y su esperanza común es

$$E_{\theta_0}[X_i] = E_{\theta_0}\left[\log\left(\frac{f(\mathbf{Y}_i;\theta)}{f(\mathbf{Y}_i;\theta_0)}\right)\right] = \nu(\theta) < 0,$$

por la parte (i). Por lo tanto, la ley de los grandes números en el Teorema 3.2.2 implica que existe un evento Ω^* con $P_{\theta_0}[\Omega^*] = 1$, tal que, para cada trayectoria $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots)$ en Ω^*,

$$\nu(\theta) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(\mathbf{Y}_i,\theta)}{f(\mathbf{Y}_i,\theta_0)}\right), \quad \omega \in \Omega^*$$

Combinando esta relación con (4.2.3) se desprende que

$$\lim_{n\to\infty} \frac{1}{n}\log\left(\frac{V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}{V_n(\theta_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}\right) = \nu(\theta), \quad \mathbf{Y} = (\mathbf{Y}_1,\mathbf{Y}_2,\ldots) \in \Omega^*$$

completando el argumento pues, como ya se ha mencionado, $P_{\theta_0}[\Omega^*] = 1$.

iii) Sea Ω^* el evento en la demostración de la parte (ii). Combinando la anterior relación desplegada con el hecho de que $\nu(\omega) < 0$, a partir de la definición de límite se obtiene que para cada $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots) \in \Omega^*$, existe un entero $N(Y,\theta)$ tal que

$$\lim_{n\to\infty} \frac{1}{n}\log\left(\frac{V_n(\theta,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}{V_n(\theta_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_n)}\right) < 0, \quad n > N(Y,\theta),$$

lo cual equivale a

$$\log\left(V_n(\theta, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)\right)$$

$$< \log\left(V_n(\theta_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)\right), \quad n > N(\mathbf{Y}, \theta),$$

o bien

$$V_n(\theta, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) < V_n(\theta_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n), \quad n > N(\mathbf{Y}, \theta)$$

Concluyendo la demostración. □

4.3 Demostración del Teorema 4.1.1

Los resultados en el Teorema 4.2.1 se utilizarán ahora para establecer la consistencia del método de verosimilitud máxima cuando el espacio de parámetros es finito.

Demostración del Teorema 4.1.1. Escriba

$$\Theta = \{\theta_0, \theta_1, \ldots, \theta_r\}, \tag{4.3.4}$$

donde θ_0 es el verdadero valor del parámetro. Para cada $i = 1, 2, \ldots, r$, por el Teorema 4.2.1 (iii) existe un evento Ω_i^* tal que

i) $P_{\theta_0}[\Omega_i^*] = 1$, y

ii) Para cada trayectoria $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots) \in \Omega_i^*$ existe un entero $N(Y, \theta_i)$ tal que

$$V_n(\theta_i, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) < V_n(\theta_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n), \quad n > N(Y, \theta_i),$$

desigualdad que muestra que

$$\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots) \in \Omega_i^* \Rightarrow \hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \neq \theta_i \quad \text{si } n > N(Y, \theta_i). \tag{4.3.5}$$

Ahora defina

$$\Omega^* = \bigcap_{i=1}^{r} \Omega_i^*$$

y note que $P_{\theta_0}[\Omega^*] = 1$. Sea

$$N(\mathbf{Y}) = \max_{i=1,2\ldots,r} N(Y, \theta_i)$$

y observe que $N(\mathbf{Y})$ es finito, pues el espacio de parámetros lo es. Seleccione ahora $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots) \in \Omega^*$ y tome $n > N(\mathbf{Y})$. En este caso $Y \in \Omega_i^*$ y $n > N(Y, \theta_i)$ para cada $i = 1, 2, \ldots, r$ y (4.3.5) implica que $\hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \neq \theta_i$. En resumen:

$$\text{Si } \mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots) \in \Omega^* \text{ y } n > N(Y),$$

$$\text{entonces } \hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \neq \theta_i, \quad i = 1, 2, \ldots, r$$

Como $\hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in \Theta$, combinando este enunciado con (4.3.4) se desprende que

$$\text{Si } Y = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots) \in \Omega^* \text{ y } n > N(Y) \text{ entonces } \hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \theta_0,$$

completando la demostración, pues como ya se ha mencionado, $P_{\theta_0}[\Omega^*] = 1$ y $N(Y)$ es finito.

□

Capítulo 5

Criterio de Consistencia

5.1 Introducción

Este trabajo trata sobre el método de verosimilitud máxima para construir estimadores puntuales, el cual es ampliamente usado en la estadística y, bajo condiciones adecuadas de regularidad que se cumplen frecuentemente en las aplicaciones, genera estimadores asintóticamente eficientes (Severini, 2001, Shao, 1999, Lehmann y Casella 1998, Azzalini, 1996). El punto de partida es una sucesión $\{X_i\}$ de objetos aleatorios independiente e idénticamente distribuidos cuya distribución común es absolutamente continua con respecto a cierta medida, y la correspondiente densidad no está completamente especificada, sino que depende de un parámetro desconocido θ. En este contexto, la propiedad asintótica más básica y deseable de un método de estimación para θ es su *consistencia*, la cual, en términos simples, requiere la convergencia de los estimadores generados hacia el verdadero valor del parámetro conforme el tamaño de la muestra crece; vea la Definición 5.2.1. En esta dirección, Bahadur (1958) mostró que el procedimiento de verosimilitud máxima no es necesariamente consistente (vea también Lehmann y Casella, 1998, p. 445-447), y este hecho proporciona la motivación para el principal problema analizado en este trabajo: *Determinar un criterio necesario y suficiente para la consistencia del método de verosimilitud máxima.*

El problema anterior se analiza, esencialmente, bajo tres tipos de supuestos presentados formalmente en las siguientes secciones: Primero, se supone que que el espacio de parámetros Θ es un espacio métrico *localmente compacto*, un requerimiento débil que se satisface, por ejemplo, cuando Θ es un (subconjunto

abierto de un) espacio Euclideano $\mathbb{R}^k$. Luego, se supone que la densidad desconocida de los objetos X_i depende continuamente, en un cierto sentido, del parámetro $\theta \in \Theta$; como es usual, la discrepancia entre dos densidades se mide en la escala logarítmica, pero no se supone que densidades asociadas a distintos parámetros tengan el mismo soporte. Finalmente, se supone que, con probabilidad uno, un estimador de verosimilitud máxima está bien definido cuando el tamaño de la muestra es suficientemente grande. En este contexto, *el principal resultado* de este trabajo, enunciado como Teorema 5.4.1 en la Sección 5.4, puede describirse como sigue:

- Una sucesión $\{\hat{\theta}_n\}$ de estimadores de verosimilitud máxima es consistente si, y sólo si, existe un conjunto compacto tal que, con probabilidad 1, dicho conjunto contiene a $\hat{\theta}_n$ para todo n suficientemente grande.

  Al aplicar este resultado al caso en que $\Theta = \mathbb{R}^k$, la siguiente caracterización se desprende de inmediato (vea el Corolario 4.1):

- Una sucesión de estimadores de verosimilitud máxima es consistente si y sólo si la sucesión es *acotada* casi seguramente.

Este último resultado es una propiedad notable del método de verosimilitud máxima, puesto que una sucesión acotada arbitraria no necesariamente converge. Los argumentos empleados a continuación dependen de dos hechos básicos, a saber,

i) la ley fuerte de los grandes números para variables cuya esperanza no es necesariamente finita, (Ash 1972, p. 277, Billingsley 1995, p. 284), y

ii) la ley cero-uno de Hewitt-Savage para eventos simétricos (Ash 1972, p. 279, Billingsley 1995, p. 496).

La organización del capítulo es la siguiente: Primeramente, en la Sección 5.2 se introduce el modelo estadístico, y se formulan las hipótesis estructurales básicas. Luego, el procedimiento de verosimilitud máxima se discute brevemente en la Sección 5.3, estableciendo una ley cero-uno para la existencia de los estimadores máximo verosímilies $\hat{\theta}_n$ para muestras grandes. Posteriormente, en la Sección 5.4 se establecen los criterios para la consistencia de $\{\hat{\theta}_n\}$; en esta parte el argumento usa una herramienta técnica establecida como Teorema 5.4.2, y la exposición concluye en la Sección 5.5 con una demostración de este resultado.

Notación 1. Para un espacio medible $(S, \mathcal{G})$ y $n = 1, 2, 3, \ldots$, S^n denota el producto cartesiano de S n veces consigo mismo, mientras que $\mathcal{G}^n$ es la σ-álgebra generada por los conjuntos $B_1 \times B_2 \times \cdots \times B_n$ con $B_i \in \mathcal{G}$ para $i = 1, 2, \ldots, n$. Similarmente, S^∞ consiste de todas las sucesiones

$$\mathbf{x} = (x_1, x_2, x_3 \ldots)$$

con $x_i \in S$ para cada i, y $\mathcal{G}^\infty$ denota la σ-álgebra generada por los cilindros $B \times \mathcal{S}^\infty$, donde $B \in \mathcal{G}^n$ para algún n. Por otro lado, para un entero positivo m, $\mathcal{P}_m$ representa la clase de todas las permutaciones de $\{1, 2, \ldots, m\}$, y para cada $\mathbf{x} \in \mathcal{S}^\infty$ y $m = 1, 2, 3, \ldots$,

$$\mathbf{x}_\tau = (x_{\tau(1)}, x_{\tau(2)}, \ldots, x_{\tau(m)}, x_{m+1}, x_{m+2}, \ldots), \quad \tau \in \mathcal{P}_m. \tag{5.1.1}$$

5.2 Modelo Estadístico

Sean $X_1, X_2, X_3, \ldots$ objetos aleatorios *i.ĩ. d.* definidos en el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$ los cuales toman valores en el espacio medible $(S, \mathcal{G})$. La distribución común de los X_i es la medida P_X definida en $\mathcal{G}$ y especificada por

$$P_X[B] = P[X_i \in B], \quad B \in \mathcal{G}. \tag{5.2.2}$$

Hipótesis 5.2.1. *Existe una medida (σ-finita) ν definida en $(S, \mathcal{G})$ tal que P_X es absolutamente continua con respecto a ν, i.e., para alguna función de densidad* $f_X : S \to [0, \infty)$,

$$P_X[B] = \int_B f_X(x)\, \nu(dx), \quad B \in \mathcal{G}. \tag{5.2.3}$$

La medida P_X —o, equivalentemente, la densidad f_X— contiene toda la información probabilística acerca de la sucesión $\{X_i\}$. Sin embargo, en la práctica f_X no es completamente conocida, y *el problema estadístico* consiste en usar los valores observados de $X_1, X_2, X_3, \ldots$ —los datos— para aproximar la densidad desconocida f_X. En adelante, se supone que se tiene información *a priori* sobre el proceso físico que genera las observaciones, el cual permite postular que f_X pertenece a cierta clase de densidades $\mathbb{F}$; la inclusión

$$f_X \in \mathbb{F} \tag{5.2.4}$$

es un *modelo estadístico*, y en este trabajo se supone que los miembros de $\mathbb{F}$ se pueden etiquetar mediante los elementos de un espacio métrico Θ, i.e.,

$$\mathbb{F} = \{f(x;\theta) : \theta \in \Theta\} \tag{5.2.5}$$

donde, para cada $\theta \in \Theta$, la función $\mathcal{G}$-medible $f(\cdot;\theta) : S \to [0,\infty)$ es una densidad con respecto a ν; en este caso (5.2.4) es un *modelo paramétrico* y Θ es el *espacio de parámetros*. *La parametrización* $\theta \mapsto f(\cdot;\theta)$ de Θ sobre $\mathbb{F}$ se supone identificable (inyectiva), en el siguiente sentido.

Hipótesis 5.2.2. *Para cada* $\theta, \theta' \in \Theta$ *con* $\theta \neq \theta'$,

$$\nu\left[x : f(x;\theta) \neq f(x;\theta')\right] > 0.$$

Bajo la condición de que $f(\cdot;\theta) = f_X(\cdot)$, la distribución común $\tilde{P}_\theta$ de los objetos X_i está dada por

$$\tilde{P}_\theta[B] = \int_B f(x;\theta)\,\nu(dx), \quad B \in \mathcal{G} \tag{5.2.6}$$

(vea (5.2.2) y(5.2.3)); la distribución de el proceso total $\mathbf{X} = (X_1, X_2, X_3, \ldots)$ se denota mediante P_θ y $E_\theta[\cdot]$ representa el correspondiente operador de valor esperado. Observe que

$$P_\theta = \tilde{P}_\theta \times \tilde{P}_\theta \times \cdots, \tag{5.2.7}$$

el producto numerable de la medida $\tilde{P}_\theta$ con ella misma. El parámetro (desconocido) $\theta^* \in \Theta$ para el cual $f_X(\cdot) = f(\cdot;\theta^*)$ es el verdadero valor del parámetro, y el problema de identificar f_X dentro de la familia $\mathbb{F}$, es el mismo que buscar θ^* dentro de Θ. En general, basándose en un número finito de observaciones $X_1, X_2, \ldots, X_n$, no es posible determinar θ^* exactamente, y los datos deben usarse para construir un estimador $\tilde{\theta}_n(X_1, X_2, \ldots, X_n)$ cuyos valores son usados como aproximaciones de θ^*.

Definición 5.2.1. *La sucesión* $\{\tilde{\theta}_n \equiv \tilde{\theta}_n(X_1, X_2, \ldots, X_n)\}$ *de estimadores de* θ *— o el método usado para construirla — es consistente si*

$$P_\theta \left[\lim_{n\to\infty} \tilde{\theta}_n = \theta \right] = 1, \quad \theta \in \Theta.$$

En las siguientes secciones se estudia el método de verosimilitud máxima, y se establecen condiciones necesarias y suficientes para su consistencia. La discusión supone que las siguientes condiciones topológicas y de continuidad se satisfacen.

Hipótesis 5.2.3. *El espacio de parámetros* Θ *es un espacio métrico localmente compacto.*

A continuación, para cada $\theta \in \Theta$ defina el soporte S_θ de la densidad $f(\cdot;\theta)$ mediante

$$S_\theta = \{x \in S\colon f(x;\theta) > 0\} \tag{5.2.8}$$

mientras que, para $\varepsilon > 0$ y $\theta_1 \in S$, la función de ε-discrepancia de la familia $\{f(\cdot;\theta)\}$ en el punto θ_1 está dada por

$$D_{\varepsilon,\theta_1}(x) = \sup_{\theta\colon d(\theta,\theta_1)<\varepsilon} \log\left(\frac{f(x,\theta)}{f(x,\theta_1)}\right) I[x \in S_\theta \cap S_{\theta_1}], \quad x \in S, \tag{5.2.9}$$

donde $d(\cdot,\cdot)$ es la métrica en Θ.

Hipótesis 5.2.4.

a) Para cada $\varepsilon > 0$ *y* $\theta, \theta_1 \in \Theta$

i) $D^+_{\varepsilon,\theta_1}(x) = \max\{0, D_{\varepsilon,\theta_1}(x)\}$ *es* $\mathcal{G}$*-medible, y*

ii) $E_\theta\left[D^+_{\varepsilon,\theta_1}(X_1)\right] \to 0$ *conforme* $\varepsilon \searrow 0$.

b) Dados θ_0 *y* $\theta_1 \in \Theta$, *si* $\nu\left[S_{\theta_0} \cap S^c_{\theta_1}\right] > 0$ *entonces existen* $B \in \mathcal{G}$ *y* $\varepsilon > 0$ *tales que*

i) $\nu[B] > 0$, *y*

ii) $B \subset S_{\theta_0} \cap S^c_\theta$ *cuando* $d(\theta_1, \theta) < \varepsilon$.

Observación 5.2.1.

a) En cierto sentido, la parte *b)* en la Hipótesis 5.2.4 garantiza que los soportes S_θ no experimentan crecimientos "repentinos" conforme $\theta \to \theta_1$. Cuando $S = \mathbb{R}^k$ y $\nu(\cdot)$ es la correspondiente medida de Lebesgue, suponga que

$$S_\theta = \{(x_1, \ldots, x_k) : a_i(\theta) \leq x_1 \leq b_i(\theta), \quad i = 1, 2, \ldots, k\}$$

para ciertas funciones $a_i(\cdot), b_i(\cdot) : \Theta \to \mathbb{R}$. En este caso, la Hipótesis 5.2.4 a) se satisface si, para cada $i = 1, 2, \ldots, k$, $a_i(\cdot)$ es semi-continua inferior y $b_i(\cdot)$ es semi-continua superior.

b) Suponga que existe un conjunto $\{\rho_1, \rho_2, \ldots\}$ el cual es denso en Θ con la siguiente propiedad, la cual es válida en todos los modelos que se usan en la práctica: **A:** Para cada $x \in S$ y $\theta \in \Theta$, puede encontrarse una subsucesión $\{\rho_{n_k}\}$ que satisface $\rho_{n_k} \to \theta$ y $f(x, \rho_{n_k}) \to f(x, \theta)$ conforme $k \to \infty$.

   En este caso el supremo en (5.2.9) puede tomarse sobre los ρ_k's que satisfacen $d(\rho_k, \theta_1) < \varepsilon$ —un conjunto numerable— y entonces $D_{\varepsilon,\theta_1}(\cdot)$ es $\mathcal{G}$-medible.

5.3 Estimación de Verosimilitud Máxima

Dado un entero positivo n y puntos $x_1, x_2, \ldots, x_n \in S$, la función de verosimilitud asociada al evento

$$[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n], \tag{5.3.10}$$

denotada por $L_n(\cdot; x_1, x_2, \ldots, x_n) : \Theta \to [0, \infty)$, se define como

$$L_n(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta), \quad \theta \in \Theta. \tag{5.3.11}$$

Note que la verosimilitud satisface

$$L_n(\cdot; x_1, x_2, \ldots, x_n) = L_n\big(\cdot; x_{\tau(1)}, x_{\tau(2)}, \ldots, x_{\tau(n)}\big), \quad \tau \in \mathcal{P}_n, \tag{5.3.12}$$

de manera que, definiendo el conjunto M_n por

$$M_n = \Big[(x_1, x_2, \ldots, x_n) \in S^n \colon L_n(\cdot; x_1, x_2, \ldots, x_n) \text{ tiene un maximizador}\Big], \tag{5.3.13}$$

se desprende que M_n es simétrico, *i.e.*,

$$(x_1, x_2, \ldots, x_n) \in M_n \iff \Big(x_{\tau(1)}, x_{\tau(2)}, \ldots, x_{\tau(n)}\Big) \in M_n, \quad \tau \in \mathcal{P}_n. \tag{5.3.14}$$

Después de observar (5.3.10), el método de verosimilitud máxima prescribe estimar θ mediante un maximizador $\hat{\theta}(x_1, x_2, \ldots, x_n)$ de $L_n(\cdot; x_1, x_2, \ldots, x_n)$ siempre y cuando tal punto exista, así que,

$$L_n\Big(\hat{\theta}(x_1, x_2, \ldots, x_n); x_1, x_2, \ldots, x_n\Big)$$

$$\geq L_n(\theta; x_1, x_2, \ldots, x_n), \quad \theta \in \Theta, \quad (x_1, x_2, \ldots, x_n) \in M_n \tag{5.3.15}$$

mientras que $\hat{\theta}_n(x_1, x_2, \ldots, x_n)$ está definido 'arbitrariamente' cuando la función de verosimilitud $L_n(\cdot; x_1, x_2, \ldots, x_n)$ no alcanza su máximo, digamos

$$\hat{\theta}(x_1, x_2, \ldots, x_n) = \theta_*, \quad (x_1, \ldots, x_n) \notin M_n \tag{5.3.16}$$

donde θ_* es un miembro fijo de Θ. Por otro lado,$\hat{\theta}_n(\cdot)$ debe ser una función medible de $(x_1, x_2, \ldots, x_n)$ para asegurar que $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$ es un estimador legítimo, y esto requiere condiciones adicionales sobre la función $(x, \theta) \mapsto f(x; \theta)$. En lugar de profundizar en temas de medibilidad, aquí se supondrá simplemente que M_n pertenece a $\mathcal{G}^n$, y que $\hat{\theta}_n(x_1, x_2, \ldots, x_n)$ es una función $\mathcal{G}^n$-medible. En estas circunstancias $\hat{\theta}_n \equiv \hat{\theta}(X_1, X_2, \ldots, X_n)$ es un estimador de verosimilitud máxima de θ basado en $X_1, X_2, \ldots, X_n$. A continuación, sea $\tilde{\theta}_n(X_1, X_2, \ldots, X_n)$ un estimador arbitrario de verosimilitud máxima, y defina

$$\hat{\theta}_n(X_1, X_2, \ldots, X_n) = \tilde{\theta}_n\Big(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\Big)$$

donde los $X_{(i)}$'s son los estadísticos de orden de la muestra $X_1, X_2, \ldots, X_n$. En este caso, a partir de (5.3.12)–(5.3.16) se desprende que $\hat{\theta}_n$ es también un estimador de verosimilitud máxima, que es simétrico, *i.e.*,

$$\hat{\theta}_n(X_1, X_2, \ldots, X_n) = \hat{\theta}_n\Big(X_{\tau(1)}, X_{\tau(2)}, \ldots, X_{\tau(n)}\Big), \quad \tau \in \mathcal{P}_n; \tag{5.3.17}$$

todos los 'buenos' estimadores tienen esta propiedad. Note ahora que, debido a que $\hat{\theta}_n(\cdot)$ está definido arbitrariamente en M_n^c, la característica esencial de una estimación de verosimilitud máxima es la desigualdad (5.3.15), así que es interesante investigar si la inclusión $(x_1, x_2, \ldots, x_n) \in M_n$ ocurre, con probabilidad 1, para n suficientemente grande; el principal resultado en esta dirección es el resultado cero-uno en el siguiente Lema 2.1, en el cual se utiliza la siguiente notación: Para $\mathbf{x} = (x_1, x_2, x_3 \ldots) \in S^\infty$ y un entero positivo n, defina

$$\mathbf{x}^n = (x_1, x_2, \ldots, x_n), \tag{5.3.18}$$

y sea M_n' el conjunto de trayectorias $\mathbf{x} \in S^\infty$ tales que, después de observar $X_1, X_2, \ldots, X_n$, la correspondiente verosimilitud L_n alcanza su máximo. Más precisamente,

$$M_n' = \left[\mathbf{x} \in S^\infty : \mathbf{x}^n \in M_n\right] = M_n \times S^\infty. \tag{5.3.19}$$

Con esta notación, $\bigcap_{m=k}^{\infty} M_m'$ consiste de todas las trayectorias $\mathbf{x} \in S^\infty$ sobre las cuales la verosimilitud correspondiente a las primeras m observaciones alcanza su máximo para todo $m \geq k$, y entonces

$$M^* = \bigcup_{k=1}^{\infty} \bigcap_{m=k}^{\infty} M_m' \tag{5.3.20}$$

es la clase de todas las trayectorias $\mathbf{x}$ para las cuales se tiene que las verosimilitudes $L_m(\cdot; \mathbf{x}^m)$ tienen maximizadores cuando m es suficientemente grande; usando la terminología en Billingsley (1995), o Shao (1999), M^* es el límite inferior de los eventos M_m'.

Lema 5.3.1. *Para cada* $\theta \in \Theta$,

$$P_\theta[M^*] = 0 \quad or \quad P_\theta[M^*] = 1.$$

Demostración del Lema 5.3.1. Sea $\theta \in \Theta$ un elemento arbitrario pero fijo, y suponga que $\theta \in \Theta$ es el verdadero valor del parámetro, de tal suerte que P_θ es la distribución de $\mathbf{X} = (X_1, X_2, \ldots)$, y entonces

$$P_\theta[M^*] = P[\mathbf{X} \in M^*]. \tag{5.3.21}$$

Sean el entero positivo m y $\tau \in \mathcal{P}_m$ arbitrarios, y observe que, con la notación en (5.1.1), para cada $t \geq m$, (5.3.13), (5.3.14) y (5.3.19) implican que $\mathbf{x} \in M'_t \iff \mathbf{x}_\tau \in M'_t$, así que

$$\mathbf{x} \in \bigcup_{k=m}^{\infty} \bigcap_{t=k}^{\infty} M'_t \iff \mathbf{x}_\tau \in \bigcup_{k=m}^{\infty} \bigcap_{t=k}^{\infty} M'_t$$

Por otro lado, observando que $\bigcap_{t=k}^{\infty} M'_t \subset \bigcap_{t=k_1}^{\infty} M'_t$ para $k \leq k_1$, se desprende a partir de (5.3.20) que

$$M^* = \bigcup_{k=m}^{\infty} \bigcap_{t=k}^{\infty} M'_t$$

por lo tanto, puesto que el entero positivo m y $\tau \in \mathcal{P}_m$ son arbitrarios, la anterior relación desplegada implica que

$$\mathbf{x} \in M^* \iff \mathbf{x}_\tau \in M^*, \quad \tau \in \mathcal{P}_m, \quad m = 1, 2, 3, \ldots, \tag{5.3.22}$$

i.e., M^* es simétrico. Recordando que los X_i's son *i.ĩ. d.* , la ley cero-uno de Hewitt-Savage para eventos simétricos implica que $P[\mathbf{X} \in M^*] = 1$ o $P[\mathbf{X} \in M] = 0$ (Ash 1972, Billingsley 1995), y entonces $P_\theta[M^*] = 0$ o $P_\theta[M^*] = 1$, por (5.3.21).□

Debido a que $\hat{\theta}_n(X_1, \ldots, X_n)$ es arbitrario cuando $(X_1, X_2, \ldots, X_n) \notin M_n$, es claro que un 'buen' comportamiento asintótico de $\{\hat{\theta}_n\}$ puede esperarse sólo

cuando $P_\theta[M^*] = 1$. Este requerimiento se satisface para todos los modelos que se consideran usualmente, y en general puede garantizarse imponiendo condiciones como continuidad de la función $\theta \mapsto f(x;\theta)$ para cada $x \in S$ y (i) compacidad de Θ, o (ii) $\sup_{\theta\in\Theta\setminus K_i} f(x;\theta) \to 0$ conforme $i \to \infty$, donde $\Theta = \bigcup_{i=1}^{\infty} K_i$ y cada conjunto K_i es compacto. En vez de dar condiciones explícitas para asegurar que $P_\theta[M^*] = 1$, se supondrá simplemente lo siguiente:

Hipótesis 5.3.1. *$M_n \in \mathcal{G}^n$ para $n = 1,2,3,\ldots$, y $P_\theta[M^*] = 1$ para cada $\theta \in \Theta$.*

5.4 Condiciones Necesarias y Suficientes Para la Consistencia

En esta sección se analiza la consistencia de una sucesión de estimadores de verosimilitud máxima. Como se muestra en el siguiente ejemplo, bajo los supuestos de este trabajo, la consistencia del método de verosimilitud máxima no está garantizada.

Ejemplo 5.4.1. Suponga que el espacio de parámetros es $\Theta = \{0,1,2,3,\ldots\}$ dotado con la métrica discreta, y sea $\varphi(x)$ una densidad sobre la recta tal que $\varphi(x) > 0$ para cada $x \in \mathbb{R} = S$. Defina $f(x;0) = \varphi(x)$ y para $k = 1,2,3,\ldots$

$$f(x;k) = \frac{\varphi(x)}{c_k} I\left[x \in [-k,k]\right], \quad \text{donde } c_k = \int_{-k}^{k} \varphi(x)\,dx.$$

En este contexto, las hipótesis de la secciones precedentes se satisfacen y, para $x_1, x_2, \ldots, x_n \in \mathbb{R}$, la función de verosimilitud $L_n(\cdot; x_1, x_2, \ldots, x_n)$ alcanza su máximo en el único punto

$$\hat{\theta}_n(x_1,\ldots,x_n) = \text{mín}\{k \in \Theta\colon |x_i| \le k,\ i = 1,2,\ldots,n\},$$

y no es difícil ver que $\hat{\theta}_n \to \infty$ con probabilidad uno con respecto a P_0. Luego, la condición $\hat{\theta}_n \to 0$ P_0-casi seguramente falla, y entonces $\{\hat{\theta}_n\}$ no es una sucesión consistente; vea la Definición 5.2.1. En Lehmann y Casella (1998, p.445) se presenta un ejemplo más sofisticado en el cual la convergencia $\hat{\theta}_n \to \theta$ P_θ casi seguramente falla *para todo* $\theta \in \Theta$.)

El siguiente teorema proporciona una condición necesaria y suficiente para que una sucesión de estimadores verosimilitud máxima tenga la propiedad de consistencia.

Teorema 5.4.1. *Suponga que las Hipótesis 5.2.1–5.2.4 así como la Hipótesis 5.3.1 se satisfacen, y sea* $\{\hat{\theta}_n \equiv \hat{\theta}_n(X_1, X_2, \ldots, X_n)\}$ *una sucesión de estimadores de verosimilitud máxima. En este caso, las siguientes afirmaciones a) y b) son equivalentes:*

a) $\{\hat{\theta}_n\}$ *es una sucesión consistente de estimadores de* θ*; vea la Definición 5.2.1.*

b) Para cada $\theta \in \Theta$*, existe un conjunto compacto* $C_\theta \subset \Theta$ *tal que, con probabilidad 1 con respecto a* P_θ*,* $\hat{\theta}_n$ *pertenece a* C_θ *para* n *suficientemente grande. Más precisamente,*

$$P_\theta \left[\bigcup_{k=1}^{\infty} \bigcap_{r=k}^{\infty} [\hat{\theta}_r \in C_\theta] \right] = 1, \quad \theta \in \Theta. \tag{5.4.23}$$

Como se muestra en el siguiente corolario, el Teorema 5.4.1 permite establecer una caracterización muy simple para modelos con $\Theta = \mathbb{R}^k$.

Corolario 5.4.1. *Suponga que* $\Theta = \mathbb{R}^k$ *y que las Hipótesis 5.2.1, 5.2.2, 5.2.4 y 5.3.1 se satisfacen. Sea* $\{\hat{\theta}_n\}$ *una sucesión de estimadores de verosimilitud máxima que satisfacen la condición de simetría (5.3.17). En este contexto,* $\{\hat{\theta}_n\}$ *es consistente si y sólo si*

$$P_\theta \left[\limsup_{n\to\infty} \|\hat{\theta}_n\| < \infty \right] = 1, \quad \theta \in \Theta. \tag{5.4.24}$$

De acuerdo a este corolario, una sucesión $\{\hat{\theta}_n\}$ de estimadores simétricos de verosimilitud máxima es consistente si, y sólo si, con probabilidad 1 con respecto a cada distribución P_θ, la sucesión $\{\hat{\theta}_n\}$ es *acotada*. Esta caracterización es una propiedad interesante del método de verosimilitud máxima, pues como es bien conocido una sucesión acotada en $\mathbb{R}^k$ no necesariamente converge (Rudin, 1968, Fulks, 1981).

Demostración del Corolario 5.4.1. Si $\{\hat{\theta}_n\}$ es consistente, (5.4.24) se desprende de la inclusiones

$$\left[\lim_{n\to\infty} \hat{\theta}_n = \theta \right] \subset \left[\limsup_{n\to\infty} \|\hat{\theta}_n\| \leq \|\theta\| \right] \subset \left[\limsup_{n\to\infty} \|\hat{\theta}_n\| < \infty \right].$$

Ahora, suponga que (5.4.24) ocurre, y sea $\theta \in \Theta$ un parámetro arbitrario pero fijo. Para cada enero k, defina el evento

$$B_k = \left[\mathbf{x} \in S^\infty : \limsup_{n\to\infty} \|\hat{\theta}_n(\mathbf{x}^n)\| \leq k\right], \tag{5.4.25}$$

así que $\bigcup_{k=1}^{\infty} B_k = \left[\limsup_{n\to\infty} \|\hat{\theta}_n\| < \infty\right]$, y la igualdad en (5.4.24) implica que existe un entero $k(\theta)$ tal que

$$P_\theta\left[B_{k(\theta)}\right] > 0. \tag{5.4.26}$$

Por otro lado, a partir de (5.4.25) y (5.3.17) se desprende que cada conjunto B_k es simétrico, *i.e.*, si $\mathbf{x} \in B_k$ entonces $\mathbf{x}_\tau$ßB_k para cada $\tau \in \mathcal{P}_m$ y $m = 1, 2, 3, \ldots$ (vea (5.1.1)), de manera que, como en la demostración del Lema 5.3.1, la ley cero-uno de Hewitt-Savage permite concluir que $P_\theta[B_k] = 0$ o $P_\theta[B_k] = 1$, y entonces

$$P_\theta\left[B_{k(\theta)}\right] = 1, \tag{5.4.27}$$

por (5.4.26). Para continuar, observe que (5.4.25) implica que si $\mathbf{x} \in B_{k(\theta)}$, entonces existe un entero $N(\mathbf{x})$ tal que

$$\|\hat{\theta}_n(\mathbf{x}^n)\| \leq k(\theta) + 1$$

para $n \geq n(\mathbf{x})$, *i.e.*,

$$\mathbf{x} \in \bigcap_{n=N(\mathbf{x})}^{\infty} \left[\|\hat{\theta}_n\| \leq k(\theta) + 1\right] \subset \bigcup_{r=1}^{\infty} \bigcap_{n=r}^{\infty} \left[\|\hat{\theta}_n\| \leq k(\theta) + 1\right],$$

esto es,

$$B_{k(\theta)} \subset \bigcup_{r=1}^{\infty} \bigcap_{n=r}^{\infty} \left[\hat{\theta}_n \in C_\theta\right],$$

donde C_θ es la bola compacta $\left\{\theta \in \mathbb{R}^k : \|\theta\| \leq k(\theta) + 1\right\}$; así,

$$P_\theta\left[\bigcup_{r=1}^{\infty} \bigcap_{n=r}^{\infty} \left[\hat{\theta}_n \in C_\theta\right]\right] = 1,$$

por (5.4.27). Puesto que $\theta \in \Theta$ es arbitrario, se sigue que $\{\hat{\theta}_n\}$ es consistente, por el Teorema 5.4.1, concluyendo el argumento. □

La demostración del Teorema 5.4.1 depende de la siguiente herramienta técnica, la cual se verificará en la siguiente sección.

Teorema 5.4.2. *Suponga que las condiciones del Teorema 5.4.1 se satisfacen, sea $\theta_0 \in \Theta$ un parámetro arbitrario pero fijo, y sea $K \subset \Theta$ un conjunto compacto tal que $\theta_0 \notin K$. En este contexto, existe $\mathcal{U}_K \in \mathcal{G}^\infty$ con las siguientes propiedades a) y b):*

a) $P_{\theta_0}[\mathcal{U}_K] = 1$;

b) Para cada $\mathbf{x} \in \mathcal{U}_K$, $\hat{\theta}_n(\mathbf{x}^n)$ no pertenece a K para n suficientemente grande. Más precisamente, existe una función $N_K : \mathcal{U}_K \to \{1, 2, 3, \ldots\}$ tal que

$$\hat{\theta}_n(\mathbf{x}^n) \notin K, \quad \mathbf{x} \in \mathcal{U}_K, \quad n \geq N_K(\mathbf{x}). \tag{5.4.28}$$

Demostración del Teorema 5.4.1. Suponga que la sucesión $\{\hat{\theta}_n\}$ es consistente. Sea $\theta \in \Theta$ un parámetro arbitrario pero fijo, y seleccione $\varepsilon_\theta > 0$ tal que la bola cerrada $C_\theta = \{\theta' \in \Theta : d(\theta', \theta) \leq \varepsilon_\theta\}$ es compacta; vea la Hipótesis 5.2.3. Observando que

$$\left[\lim_{n\to\infty} \hat{\theta}_n = \theta\right] \subset \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left[d(\hat{\theta}_n, \theta) \leq \varepsilon_\theta\right] = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left[\hat{\theta}_n \in C_\theta\right],$$

la consistencia de $\{\hat{\theta}_n\}$ implica que (5.4.23) ocurre.

Suponga que (5.4.23) ocurre, donde cada conjunto C_θ es compacto. En este caso sea $\theta_0 \in \Theta$ un parámetro arbitrario, y note los siguientes hechos a)–c):

a) Si $\mathbf{x} \in \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left[\hat{\theta}_n \in C_{\theta_0}\right]$, entonces existe un entero $k(\mathbf{x})$ tal que

$$\mathbf{x} \in \bigcap_{n=k(\mathbf{x})}^{\infty} \left[\hat{\theta}_n \in C_{\theta_0}\right],$$

i.e.,

$$\hat{\theta}_n(\mathbf{x}^n) \in C_{\theta_0}, \quad n \geq k(\mathbf{x}).$$

b) Dado $\varepsilon > 0$, sea $B(\theta_0, \varepsilon) = \{\theta' \in \Theta\colon d(\theta', \theta_0) < \varepsilon\}$ la bola abierta con centro θ_0 y radio $\varepsilon > 0$, y defina el conjunto $K = C_{\theta_0} \cap B(\theta_0, \varepsilon)^c$, de manera que K es compacto, $\theta_0 \notin K$, y

$$C_{\theta_0} = K \cup \left(C_{\theta_0} \cap B(\theta_0, \varepsilon)\right).$$

c) Puesto que $\theta_0 \notin K$, por el Teorema 5.4.2 existe $\mathcal{U}_K \in \mathcal{B}(S^\infty)$ con $P_{\theta_0}[\mathcal{U}_K] = 1$, así como una función $N_K(\cdot)\colon \mathcal{U}_K \to \{1, 2, 3, \ldots\}$ que satisface que, para cada $\mathbf{x} \in \mathcal{U}_K$,

$$\hat{\theta}_n(\mathbf{x}^n) \notin K, \quad n \geq N_K(\mathbf{x});$$

vea (5.4.28). Definiendo $N(\mathbf{x}) = \text{máx}\{k(\mathbf{x}), N_K(\mathbf{x})\}$, combinando a)–c) se desprende que

$$\mathbf{x} \in \mathcal{U}_k \cap \left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} [\hat{\theta}_n \in C_{\theta_0}]\right) \Longrightarrow \hat{\theta}_n(\mathbf{x}) \in C_{\theta_0} \cap B(\theta_0, \varepsilon), \quad n \geq N(\mathbf{x})$$

$$\Longrightarrow \mathbf{x} \in \left[\hat{\theta}_n \in B(\theta_0, \varepsilon)\right], \quad n \geq N(\mathbf{x})$$

$$\Longrightarrow \mathbf{x} \in \bigcap_{n=N(\mathbf{x})}^{\infty} \left[d(\hat{\theta}_n, \theta_0) < \varepsilon\right],$$

así que

$$\mathcal{U}_k \cap \left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} [\hat{\theta}_n \in C_{\theta_0}]\right) \subset \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left[d(\hat{\theta}_n, \theta_0) < \varepsilon\right],$$

y entonces

$$P_{\theta_0}\left[\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left[d(\hat{\theta}_n, \theta_0) < \varepsilon\right]\right] = 1.$$

Puesto que esta última igualdad se satisface para cada $\varepsilon > 0$, se sigue que $P_{\theta_0}\left[\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right] = 1$, y entonces $\{\hat{\theta}_n\}$ es una sucesión consistente de estimadores de θ, pues $\theta_0 \in \Theta$ es arbitrario.

□

5.5 Demostración del Teorema 5.4.2

En el resto del trabajo, $\theta_0 \in \Theta$ es arbitrario pero fijo, se supone que las hipótesis en las anteriores Secciones 5.2 y 5.3 se satisfacen aún sin mención adicional, y $\{\hat{\theta}_n\}$ es una sucesión dada de estimadores de verosimilitud máxima de θ. Como es usual, en lugar de analizar las funciones $L_n(\cdot; \mathbf{x}^n)$ directamente, es conveniente considerar sus logaritmos normalizados

$$\mathcal{L}_n(\theta; \mathbf{x}^n) = \frac{1}{n}\log\left(L_n(\theta; \mathbf{x}^n)\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(f(x_i;\theta)\right), \quad \theta \in \Theta, \tag{5.5.29}$$

donde se adopta la convención $\log(0) = -\infty$; puesto que $\log(\cdot)$ es estrictamente creciente en $[0, \infty)$, (5.3.15) equivale a

$$\mathcal{L}_n\left(\hat{\theta}(\mathbf{x}^n); \mathbf{x}^n\right) \geq \mathcal{L}_n(\theta; \mathbf{x}^n), \quad \theta \in \Theta, \quad \mathbf{x}^n \in M_n. \tag{5.5.30}$$

La demostración del Teorema 5.4.2 se basa en el siguiente resultado.

Teorema 5.5.1. *Dado $\theta_1 \in \Theta$ con $\theta_1 \neq \theta_0$, existen números positivos $\varepsilon(\theta_1)$ y $\Delta(\theta_1)$, así como un evento $\mathcal{U}_{\theta_1} \in \mathcal{G}^\infty$ y una función $N_{\theta_1} : \mathcal{U}_{\theta_1} \to \{1, 2, 3, \ldots\}$ tal que las siguientes propiedades (a) y (b) se satisfacen:*

a) $\mathcal{U}_{\theta_1} \subset S^\infty_{\theta_0}$ y $P_{\theta_0}\left[\mathcal{U}_{\theta_1}\right] = 1$;

b) Para cada $\mathbf{x} \in \mathcal{U}_{\theta_1}$

$$\mathcal{L}_n\left(\theta; \mathbf{x}^n\right) \leq \mathcal{L}_n\left(\theta_0; \mathbf{x}^n\right) - \Delta(\theta_1), \quad \text{si} \quad n \geq N_{\theta_1}(\mathbf{x}) \quad \text{y} \quad d(\theta, \theta_1) < \varepsilon(\theta_1).$$

El argumento usado a continuación para establecer este teorema depende de las siguientes consecuencias de la desigualdad de Jensen.

Lema 5.5.1. *Si* $\theta_1 \in \Theta \setminus \{\theta_0\}$ *y* $\nu\left(S_{\theta_0} \cap S_{\theta_1}^c\right) = 0$, *entonces las siguientes afirmaciones a) y b) ocurren.*

a) La siguiente desigualdad se satisface:

$$E_{\theta_0}\left[\log\left(\frac{f(X_1;\theta_1)}{f(X_1;\theta_0)}\right)\right] < 0, \tag{5.5.31}$$

donde la esperanza puede ser $-\infty$.

b) Existen números positivos $\varepsilon(\theta_1)$ *y* $\Delta(\theta_1)$ *tales que*

$$E_{\theta_0}\left[D^+_{\varepsilon,\theta_1}(X_1) + \log\left(\frac{f(X_1;\theta_1)}{f(X_1;\theta_0)}\right)\right] \leq -2\Delta(\theta_1), \quad si \quad 0 < \varepsilon \leq \varepsilon(\theta_1).$$

Demostración del Lema 5.5.1.

a) Debido a que $\nu(S_{\theta_0} \cap S_{\theta_1}^c) = 0$, via (5.2.6) y (5.2.8) se sigue que

$$\begin{aligned} 1 &= \tilde{P}_{\theta_0}\left[S_{\theta_0}\right] \\ &= \int_{S_{\theta_0}} f(x;\theta_0)\,\nu(dx) \\ &= \int_{S_{\theta_0}\cap S_{\theta_1}} f(x;\theta_0)\,\nu(dx) \qquad (5.5.32) \\ &= \tilde{P}_{\theta_0}\left[S_{\theta_0} \cap S_{\theta_1}\right], \end{aligned}$$

y

$$E_{\theta_0}\left[\log\left(\frac{f(X_1;\theta_1)}{f(X_1;\theta_0)}\right)\right] = \int_{S_{\theta_0}\cap S_{\theta_1}} \log\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right) f(x;\theta_0)\,\nu(dx) \tag{5.5.33}$$

Suponga ahora que $\frac{f(x;\theta_1)}{f(x;\theta_0)}$ no es constante ν-casi seguramente en $S_{\theta_0} \cap S_{\theta_1}$. En este caso, usando la concavidad estricta de $\log(\cdot)$, la relación (5.5.31) se desprende de las dos anteriores relaciones desplegadas via la desigualdad de Jensen. Para concluir, es suficiente mostrar que (5.5.31) se satisface cuando, para algún $c \in [0, \infty)$,

$$\frac{f(x;\theta_1)}{f(x;\theta_0)} = c \quad \nu\text{-casi en todas partes en} S_{\theta_0} \cap S_{\theta_1}. \tag{5.5.34}$$

En este caso, puesto que $f(\cdot;\theta_1)$ es una densidad,

$$\begin{aligned} 1 &= \int_{S_{\theta_1}} f(x;\theta_1)\nu(dx) \\ &\geq \int_{S_{\theta_0}\cap S_{\theta_1}} f(x;\theta_1)\nu(dx) \\ &= c\int_{S_{\theta_0}\cap S_{\theta_1}} f(x;\theta_0)\nu(dx) \\ &= c, \end{aligned}$$

donde (5.5.32) se usó para establecer la última igualdad. Por otro lado, a partir de (5.5.33) es claro que (5.5.31) se desprenderá si es posible demostrar que $c < 1$, desigualdad que será establecida a continuación. Suponga que $c = 1$, así que la anterior relación desplegada implica que $1 = \int_{S_{\theta_0}\cap S_{\theta_1}} f(x;\theta_1)\nu(dx) = \int_{S_{\theta_1}} f(x;\theta_1)\nu(dx)$; debido a que $f(\cdot;\theta_1)$ es positiva en S_{θ_1}, se tiene que

$$\nu\left[S_{\theta_1} \cap S_{\theta_0^c}\right] = 0$$

y, más aún, via (5.5.34), la igualdad $c = 1$ implica que

$$\nu\Big[\,[x\colon f(x;\theta_1) \neq f(x;\theta_0)] \cap (S_{\theta_0} \cap S_{\theta_1})\big]\,\Big] = 0.$$

Observando que

$$[x\colon f(x;\theta_1) \neq f(x;\theta_0)] \subset S_{\theta_0} \cup S_{\theta_1},$$

la condición $\nu\left[S_{\theta_0} \cap S_{\theta_1}^c\right] = 0$ y las dos últimas relaciones desplegadas implican que $\nu\,[x\colon f(x;\theta_1) \neq f(x;\theta_0)] = 0$, lo cual contradice el supuesto de identificabilidad en la Hipótesis 5.2.2, puesto que $\theta_1 \neq \theta_0$. Por lo tanto, $c < 1$ completando la demostración de la parte a).

b) Usando la parte a), seleccione $\Delta(\theta_1) > 0$ tal que

$$E_{\theta_0}\left[\log\left(\frac{f(X_1;\theta_1)}{f(X_1;\theta_0)}\right)\right] < -3\Delta(\theta_1).$$

Por la Hipótesis 5.2.4 b), existe $\varepsilon(\theta_1) > 0$ tal que $E_{\theta_0}\left[D^+_{\varepsilon,\theta_1}(X_1)\right] < \Delta(\theta_1)$ para $\varepsilon \in (0, \varepsilon(\theta_1)]$, y la parte b) se desprende de inmediato. □

La siguiente consecuencia de (5.2.9) será útil.

Lema 5.5.2. *Sea $\theta_1 \in \Theta \setminus \{\theta_0\}$ un parámetro arbitrario. Para cada $x \in S_{\theta_0} \cap S_{\theta_1}$ y $\varepsilon > 0$*

$$\log\left(f(x;\theta)\right) \leq D^+_{\varepsilon,\theta_1}(x) + \log\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right) + \log\left(f(x;\theta_0)\right) \quad si \quad d(\theta,\theta_1) < \varepsilon. \tag{5.5.35}$$

Demostración del Lema 5.5.2. Sea $x \in S_{\theta_0} \cap S_{\theta_1}$ un elemento arbitrario, de manera que $f(x;\theta_1)f(x;\theta_0) > 0$. Primeramente, note que (5.5.35) es válida cuando $x \notin S_\theta$, puesto que el lado izquierdo es $-\infty$. A continuación, suponga que $x \in S_\theta$, de modo que $I[x \in S_\theta \cap S_{\theta_1}] = 1$. A partir de

$$f(x;\theta) = \left[\frac{f(x;\theta)}{f(x;\theta_1)}\right]\left[\frac{f(x;\theta_1)}{f(x;\theta_0)}\right] f(x;\theta_0)$$

se desprende que

$$\log\left(f(x;\theta)\right) = \log\left(\frac{f(x;\theta)}{f(x;\theta_1)}\right) + \log\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right) + \log\left(f(x;\theta_0)\right)$$

$$= \log\left(\frac{f(x;\theta)}{f(x;\theta_1)}\right) I[x \in S_\theta \cap S_{\theta_1}]$$

$$+ \log\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right) + \log\left(f(x;\theta_0)\right)$$

y, via (5.2.9), se obtiene (5.5.35). □

Demostración del Teorema 5.5.1. El argumento se divide en dos casos, de acuerdo al valor de $\nu[S_{\theta_0} \cap S_{\theta_1}^c]$.

Caso 1: $\nu[S_{\theta_0} \cap S_{\theta_1}^c] = 0$.

Para empezar, observe que (5.2.6) y (5.2.8) implican que

$$1 = \tilde{P}_{\theta_0}[S_{\theta_0}] = \tilde{P}_{\theta_0}[S_{\theta_0} \cap S_{\theta_1}],$$

de modo que

$$P_{\theta_0}\left[\left(S_{\theta_0} \cap S_{\theta_1}\right)^\infty\right] = 1; \tag{5.5.36}$$

vea (5.2.7). A continuación, sean $\Delta(\theta_1)$ y $\varepsilon(\theta_1)$ como en el Lema5.5.1 b), y defina M como la clase de todas las trayectorias $\mathbf{x} = (x_1, x_2, x_3, \ldots) \in S^\infty$ tales que

$$\lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{n} D^+_{\varepsilon(\theta_1),\theta_1}(x_i) + \sum_{i=1}^{n} \log\left(\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}\right)\right)$$

$$= E_{\theta_0}\left[D^+_{\varepsilon(\theta_1),\theta_1}(X_1) + \log\left(\frac{f(X_1;\theta_1)}{f(X_1;\theta_0)}\right)\right]$$

$$\leq -2\Delta(\theta_1).$$

Por lo tanto

$$P_{\theta_0}[M] = 1, \tag{5.5.37}$$

por la ley de los grandes números, y existe $N_{\theta_1}(\cdot)\colon M \to \{1,2,3,\ldots\}$ tal que

$$\frac{1}{n}\left(\sum_{i=1}^{n} D^+_{\varepsilon(\theta_1),\theta_1}(x_i) + \sum_{i=1}^{n} \log\left(\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}\right)\right)$$

$$< -\Delta(\theta_1), \quad \mathbf{x} \in M, \quad n \geq N_{\theta_1}(\mathbf{x}). \tag{5.5.38}$$

Por otro lado, para $\mathbf{x} \in \left(S_{\theta_0} \cap S_{\theta_1}\right)^\infty$, se sigue que $x_i \in S_{\theta_0} \cap S_{\theta_1}$ para cada i, así que

$$\log\left(f(x_i;\theta)\right) \leq D^{+}_{\varepsilon(\theta_1),\theta_1}(x_i)$$

$$+\log\left(\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}\right)+\log\left(f(x_i;\theta_0)\right) \quad \text{si} \quad d(\theta,\theta_1)<\varepsilon(\theta_1),$$

por el Lema 5.5.2, y entonces (vea (5.5.29))

$$\mathcal{L}_n(\theta;\mathbf{x}^n) \leq \frac{1}{n}\left(\sum_{i=1}^{n} D^{+}_{\varepsilon(\theta_1),\theta_1}(x_i)+\sum_{i=1}^{n}\log\left(\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}\right)\right)+\mathcal{L}_n(\theta_0;\mathbf{x}^n)$$

$$\text{si } \mathbf{x}\in\left(S_{\theta_0}\cap S_{\theta_1}\right)^{\infty} \text{ y } d(\theta,\theta_1)<\varepsilon(\theta_1)$$

Definiendo $\mathcal{U}_{\theta_1}=\left(S_{\theta_0}\cap S_{\theta_1}\right)^{\infty}\cap M$, la parte a) en el Teorema 5.5.1 se desprende de (5.5.36) y (5.5.37), mientras que la parte b) se obtiene combinando la anterior relación desplegada y (5.5.38).

Caso 2: $\nu[S_{\theta_0}\cap S^{c}_{\theta_1}]>0$.

En este contexto, la Hipótesis 5.2.4 a) permite seleccionar $\varepsilon(\theta_1)>0$ y $B\in\mathcal{G}$ tales que $\nu[B]>0$ y

$$B\subset S_{\theta_0}\cap S^{c}_{\theta} \text{ cuando } d(\theta,\theta_1)<\varepsilon(\theta_1). \tag{5.5.39}$$

Puesto que la densidad $f(\cdot;\theta_0)$ es positiva en S_{θ_0}, se sigue que

$$\tilde{P}_{\theta_0}[B]=\int_B f(x;\theta_0)\,\nu(dx)>0$$

definiendo

$$\mathcal{U}_{\theta_1}=\left[\mathbf{x}\in S^{\infty}_{\theta_0}\colon \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I[x_i\in B]=\tilde{P}_{\theta_0}[B]\right],$$

la ley de los grandes números implica que $P_{\theta_0}[\mathcal{U}_{\theta_1}]=1$, así que $\mathcal{U}_{\theta_1}$ satisface las conclusiones en la parte a) del Teorema 5.5.1. Para finalizar, se verificará la parte b) del Teorema 5.5.1. Defina $N_{\theta_1}:\mathcal{U}_{\theta_1}\to\{1,2,3,\ldots\}$ mediante

$$N_{\theta_1}(\mathbf{x}) = \text{mín}\{i : \ x_i \in B\}, \quad \mathbf{x} \in \mathcal{U}_{\theta_1}$$

y note que $N_{\theta_1}(\cdot)$ es finito, puesto que $\tilde{P}_{\theta_0}[B] > 0$. Para $\mathbf{x} \in \mathcal{U}_{\theta_1}$ y $n > N_{\theta_1}(\mathbf{x})$ existe un entero positivo $i < n$ tal que $x_i \in B$; de hecho, $i = N_{\theta_1}(\mathbf{x})$ es tal entero. En este caso, $f(x_i;\theta) = 0$ si $d(\theta,\theta_1) < \varepsilon(\theta_1)$, por (5.5.39) y (5.2.8), y entonces

$$\mathcal{L}_n(\theta;\mathbf{x}^n) = -\infty, \quad n > N_{\theta_1}(\mathbf{x}), \quad d(\theta,\theta_1) < \varepsilon(\theta_1), \quad \mathbf{x} \in \mathcal{U}_{\theta_1};$$

vea (5.5.29) y recuerde la convención $\log(0) = -\infty$. Por otro lado, observe que $x_i \in S_{\theta_0}$ para todo i cuando $\mathbf{x} \in \mathcal{U}_{\theta_1}$, y en este caso $f(x_i;\theta_0) > 0$, así que $\mathcal{L}_n(\theta_0;\mathbf{x}^n) \in \mathbb{R}$, por (5.5.29), y entonces

$$\mathcal{L}_n(\theta;\mathbf{x}^n) < \mathcal{L}_n(\theta_0;\mathbf{x}^n) - 1, \quad n > N_{\theta_1}(\mathbf{x}), \quad d(\theta,\theta_1) < \varepsilon(\theta_1), \quad \mathbf{x} \in \mathcal{U}_{\theta_1},$$

completando el argumento. □

Demostración del Teorema 5.4.2. Sea K un subconjunto compacto arbitrario de Θ tal que $\theta_0 \in \Theta \cap K^c$. Para cada $\theta' \in K$, el Teorema 5.5.1 implica la existencia de números positivos $\varepsilon(\theta')$ y $\Delta(\theta')$, así como un evento $\mathcal{U}_{\theta'} \in \mathcal{G}^\infty$ y una función $N_{\theta'} : \mathcal{U}_{\theta'} \to \{1,2,3,\ldots\}$ tal que

$$\mathcal{U}_{\theta'} \subset S^\infty_{\theta_0}, \quad P_{\theta_0}[\mathcal{U}_{\theta'}] = 1, \tag{5.5.40}$$

y

$$\mathcal{L}_n(\theta;\mathbf{x}^n) \leq \mathcal{L}_n(\theta_0;\mathbf{x}^n) - \Delta(\theta'), \quad \mathbf{x} \in \mathcal{U}_{\theta'}, \quad n > N_{\theta'}(\mathbf{x}), \quad \theta \in B(\theta',\varepsilon(\theta')) \tag{5.5.41}$$

donde, como antes, $B(\theta',\varepsilon(\theta'))$ es la bola abierta Θ con centro θ' y radio $\varepsilon(\theta')$. Observando que $K \subset \bigcup_{\theta' \in K} B(\theta',\varepsilon(\theta'))$, la compacidad de K implica que, para cierto conjunto finito $\{\theta_1,\theta_2,\ldots,\theta_r\} \subset K$

$$K \subset \bigcup_{i=1}^{r} B(\theta_i,\varepsilon(\theta_i)). \tag{5.5.42}$$

Definiendo

$$\tilde{\mathcal{U}}_K = \bigcap_{i=1}^{r} \mathcal{U}_{\theta_i},$$

$$\tilde{N}_K(\mathbf{x}) = \text{máx}\left\{N_{\theta_1}(\mathbf{x}), N_{\theta_2}(\mathbf{x}), \ldots, N_{\theta_r}(\mathbf{x})\right\}$$

y

$$\Delta_K = \text{mín}\left\{\Delta(\theta_1), \Delta(\theta_2), \ldots, \Delta(\theta_r)\right\} > 0,$$

(5.5.40) implica que

$$\tilde{\mathcal{U}}_K \subset S^{\infty}_{\theta_0}, \quad \text{y} \quad P_{\theta_0}\left[\tilde{\mathcal{U}}_K\right] = 1, \tag{5.5.43}$$

mientras que (5.5.41) y (5.5.42) implican que

$$\mathcal{L}_n\left(\theta; \mathbf{x}^n\right) \leq \mathcal{L}_n\left(\theta_0; \mathbf{x}^n\right) - \Delta_K, \quad \mathbf{x} \in \tilde{\mathcal{U}}_K, \quad n > \tilde{N}_K(\mathbf{x}), \quad \theta \in K. \tag{5.5.44}$$

Por otro lado, a partir de la definición de M^*, se desprende que existe una función $N\colon M^* \to \{1,2,3,\ldots\}$ tal que si $\mathbf{x} \in M^*$ entonces $\mathbf{x}^n \in M_n$ para $n > N(\mathbf{x})$; vea (5.3.19) y (5.3.20). Combinando este último hecho con (5.5.30), se sigue que

$$\mathcal{L}_n\left(\hat{\theta}(\mathbf{x}^n); \mathbf{x}^n\right) \geq \mathcal{L}_n\left(\theta; \mathbf{x}^n\right), \quad \mathbf{x} \in M^*, \quad n > N(\mathbf{x}), \quad \theta \in \Theta. \tag{5.5.45}$$

Para concluir, defina $\mathcal{U}_K = \tilde{\mathcal{U}}_K \cap M^*$. Con esta notación, la Hipótesis 5.3.1 y (5.5.43) implican que $\mathcal{U}_K \subset S^{\infty}_{\theta_0}$ y $P_{\theta_0}[\mathcal{U}_K] = 1$, mientras que definiendo $N_K(\mathbf{x}) = \text{máx}\left\{\tilde{N}_K(\mathbf{x}), N(\mathbf{x})\right\}$ para $\mathbf{x} \in \mathcal{U}_K$, y usando que $\Delta_K > 0$, las relaciones (5.5.44) y (5.5.45) implican que $\hat{\theta}_n(\mathbf{x}^n) \notin K$ para $\mathbf{x} \in \mathcal{U}_K$ y $n > N_K(\mathbf{x})$, completando la demostración. □

LITERATURA CITADA

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

[2] Bahadur, R. R., (1958). Examples of inconsitency of maximum likelihood estimates, *Sankhya*, **18**, 211–224.

[3] Billingsley, P., (1995). *Probabilty and Measure, Wiley*, New York.

[4] Casella, G., Berger R., (2001). *Statistical Inference, Duxbury Press*, New York.

[5] Dudewicz E., Mishra N., (1989). *Modern Mathematical Statistics, WIley*, New York.

[6] Dudley, R. M., (2002). *Real Analysis and Probability, Cambridge University Press*, Boston.

[7] Dugundji J., (1970). *Topology, Allyn & Bacon*, Boston.

[8] Fulks, W., (1981). *Cálculo Avanzado, Limusa*, México D.F.

[9] Greene, W. H., (2003). *Econometric Analysis, Prentice–Hall*, New York.

[10] Griffiths, W. E., Carter-Hill R., Judge G. G., (1997). *Learning and Practicing Econometrics, Wiley & Sons*, New York.

[11] Lehmann E. L., Casella G., (1998). *Theory of Point Estimation, Second Edition, Springer*, New York.

[12] Lehmann E. L., (2001). *Testing Statistical Hypotheses, Wiley* , New York.

[13] Mood, A. M., Graybill F. A., Boes D. C., (1987). *Introduction to the Theory of Statistics, McGraw-Hill*, New York

[14] Hardin J. W., (2002). The robust variance estimator for two-stage models, *The Stata Journal*, **3**, pp. 253–266.

[15] Munkres. J. R., (1989). *Topology, Prentice-Hall*, New York.

[16] Newey, W., McFadden, D., (1993). Estimation in large samples. In D. McFadden and R. Engler (eds), *Handbook of Econometrics*, Vol. 4., North-Holland, Amsterdam.

[17] Murphy K., Topel R., (1985). Estimation and Inference in Two–Step Econometric Models, *Journal of Business Economics and Statistics*, **3**, October, 370-379.

[18] Randles R. H., Wolfe, D. A., (1979). *Introduction to the theory of nonparametric statistics, Wiley*, New York.

[19] Rao C. R., (2002). *Linear Statistical Inference and Its Application, Wiley*, New York.

[20] Rudin W., (1968). *Real and Complex Analysis, McGraw-Hill*, New York

[21] Serfling, R. J., (1988). *Approximation Theorems of Mathematical Statistics, Wiley*, New York.

[22] Severini T. A., (2001). *Likelihood Methods in Statistics, Oxford University Press*, London.