# CONSISTENCIA DE ESTIMADORES DE VEROSIMILITUD MAXIMA BAJO SUPUESTOS DE CONTINUIDAD

JULIO SAUCEDO ZUL

TESIS

Presentada como Requisito Parcial
para Obtener el Grado de:

Maestro en Ciencias

en Estadística Experimental

Universidad Autónoma Agraria
Antonio Narro

PROGRAMA DE GRADUADOS



Buenavista, Saltillo, Coahuila, México.

Diciembre de 2004

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

SUBDIRECCIÓN DE POSTGRADO

# CONSISTENCIA DE ESTIMADORES DE VEROSIMILITUD MÁXIMA BAJO SUPUESTOS DE CONTINUIDAD

TESIS

POR

JULIO SAUCEDO ZUL

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

MAESTRO EN CIENCIAS EN
ESTADÍSTICA EXPERIMENTAL

COMITÉ PARTICULAR

Asesor Principal: _____
Dr. Rolando Cavazos Cadena

Asesor: _____
MC. Félix de Jesús Sánchez Pérez

Asesor: _____
MC. Luis Rodríguez Gutiérrez

Asesor: _____
MC. Manuel Torres Gomar

_____
Dr. Jerónimo Landeros Flores
Subdirector de Postgrado

Buenavista , Saltillo, Coahuila. Diciembre 2004

A QUIEN SIEMPRE ESTA EN MI MENTE

JESÚS SAUCEDO ZUL †

# Agradecimientos

gradezco a mi esposa Elizabeth Flores y nuestros hijos Jesús y Emiliano por su apoyo y delidad, a mis padres Faustino Jesús y Hortensia, y a mis hermanos por su grandiosa nistad. Sin duda alguna una gran persona que influyo en mi para seguir trabajando onestamente, con su amistad, enseñanza, paciencia y gracias a esto terminar con el este abajo, mi asesor Dr. Rolando Cavazos Cadena, muchas gracias.

gradezco también a todos mis maestros que me dieron cursos durante el periodo de mi aestría así como a mis compañeros, Roger, Claudia , Edgar y Fernando.

gradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo orgado a través de la beca nacional para estudios de maestría.

COMPENDIO

# CONSISTENCIA DE ESTIMADORES DE VEROSIMILITUD MÁXIMA BAJO SUPUESTOS DE CONTINUIDAD

POR:

JULIO SAUCEDO ZUL

MAESTRIA EN CIENCIAS EN
ESTADÍSTICA EXPERIMENTAL

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

Buenavista, Saltillo, Coahuila. Diciembre 2004

Dr. Rolando Cavazos Cadena. Asesor

Palabras Clave: Consistencia de estimadores, Función cóncova, Ley fuerte de los grandes números, Convergencia casi segura, Conjunto cerrado y acotado, Conjunto abierto.

ABSTRACT

# ESTIMATORS CONSISTENCE OF MAXIMUM LIKELIHOOD HUSHED SUPPOSED OF CONTINUITY

BY

JULIO SAUCEDO ZUL

MASTER OF SCIENCE
EXPERIMENTAL STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

Buenavista, Saltillo, Coahuila. Diciembre 2004

Dr. Rolando Cavazos Cadena. Advisor

Key Words: Estimators consistence, Function concave,
Law of large numbers, Odds on convergence,
Closed set, open set.

# Indice de Contenido

**1. Presentación**

1.1 La Idea Principal 1

1.2 Familias de Densidades 2

1.3 El Problema 4

1.4 Los Instrumentos 5

1.5 La Organización 5

**2. Consistencia**

2.1 Introducción 7

2.2 Ley de los Grandes Números 9

2.3 Desigualdad de Jensen 13

2.4 El Método de Verosimilitud Máxima: Motivación 17

**3. Estimación en el Caso de Espacio de Parámetros Compacto**

3.1 Introducción 23

3.2 Supuestos de Continuidad–Compacidad 24

3.3 El Resultado Principal 30

3.4 Demostración del Teorema 3.3.1 33

**4. El Argumento Clásico**

4.1 Introducción 37

4.2 Resultados Auxiliares 39

4.3 Cotas Para la ecuación de Verosimilitud 42

4.4 Raíces de las Cotas Cuadráticas 45

4.5 Demostración del Teorema 4.1.1 48

**Literatura Citada** **51**

# Capítulo 1

# Presentación

## 1.1 La Idea Principal

Este trabajo trata sobre una idea básica en el problema de estimación estadística, el cual, en términos generales, puede describirse como sigue: Después de observar vectores aleatorios $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$, se desea utilizarlos para obtener información acerca de una cantidad desconocida $\psi$. Una manera de abordar el problema es construir, mediante algún método, una aproximación

$$\hat{\psi}_n = \hat{\psi}_n(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$$

para $\psi$ en términos de los datos observados, la cual recibe el nombre de estimador de $\psi$. El método que genera a los estimadores $\hat{\psi}_n$, o la sucesión misma de estimadores $\{\hat{\psi}_n\}$, se denomina consistente si, conforme el número de datos observados $n$ se incrementa, los valores de $\hat{\psi}_n$ convergen, en algún sentido, al valor desonocido $\psi$ (Casella y Berger, 2001, Mood *et. al.* 1987, Dudewicz y Mishra 1988). 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 recompenzado con aproximaciones cada vez más cercanas al valor desconocido $\psi$; 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.*

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 $\mu$, entonces $\hat{\mu}_n = (Y_1 + \ldots Y_n)/n$—la media muestral basada en los $n$ datos—converge hacia $\mu$. Hay varias verisones 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 $\mu$ (Ash 1987, Dudley 2002). Este resultado fundamental, que se encuentra ligado a la propia interpretación frecuencial de probabilidad, es el instrumento básico para estudiar la consistencia de estimadores.

## 1.2 Familias de Densidades

Cuando se afirma que el problema de estimación consiste en usar los datos observados para obtener aproximaciones a alguna cantidad desconocida $\psi$, la primera pregunta que surge es ¿porqué no se conoce a $\psi$? Como punto de partida para analizar este cuestionamiento, recuerde que las cantidades susceptibles de un estudio estadístico son aquellas que dependen de la distribución de los datos. Por lo tanto, el hecho de no conocer a $\psi$ implica que la distribución de los datos no está completamente especificada, que algún aspecto de ella no es conocido. En adelante se denotará mediante $\theta$ a esos 'aspectos' desconocidos de la distribución de los datos. Denote a los datos mediante $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, donde los vectores aleatorios $\mathbf{Y}_i$ están definidos

en un mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$. La probabilidad del evento $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in A$ es $P[(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in A]$ pero, en general, ésta no se calcula usando la medida de probabilidad $P$; más bien, se emplea en el cómputo la distribución del vector $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Suponiendo que $\mathbf{Y}$ tiene densidad $\Delta(\mathbf{y})$ , se calcula

$$P[(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in A] = \int_A \Delta(\mathbf{y})\, d\mathbf{y}, \tag{2.1}$$

Cualquiera de los términos en esta igualdad es la distribución de $\mathbf{Y}$ evaluada en $A$, y si $\Delta(\mathbf{y})$ se conociera con precisión, $P[(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in A]$ estaría completamente determinada. Sin embargo, al no conocer el valor de $\psi$, el cual depende de la distribución de $\mathbf{Y}$, esto es de $\Delta(\mathbf{y})$, se tiene que la densidad $\Delta(\mathbf{y})$ no está completamente determinada:

$$\Delta(\mathbf{y}) = f(\mathbf{y}; \theta)$$

donde, como ya se ha mencionado, $\theta$ representa los aspectos desconocidos de la distribución, y se supondrá que $\theta$ es un vector con componentes reales. En este caso, (2.1) se convierte en

$$P[(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in A] = \int_A f(\mathbf{y}; \theta)\, d\mathbf{y}. \tag{2.2}$$

Conocer $\theta$ implica conocer completamente la distribución de $\mathbf{Y}$ y entonces determinar $\psi$ de manera única. Por lo tanto, las cantidades desconocidas de interés $\psi$ son funciones de $\theta$, esto es, $\psi = \psi(\theta)$. En resumen: el problema de estimación se presenta cuando la densidad del vector de datos $\mathbf{Y}$ depende de algún vector desconocido $\theta$ y, bajo el supuesto de que $\mathbf{Y}$ tiene densidad, esto significa que lo que se conoce de la verdadera densidad es que pertenece a la familia $\{f(\cdot; \theta)\}$, pero no se conoce de forma precisa cual valor $\theta_0$ de $\theta$ produce la verdadera densidad de $\mathbf{Y}$. Como es usual en el análisis estadístico, se supondrá que paramentros $\theta$ distintos inducen densidades diferentes, y entonces determinar cual de las densidades dentro de la familia $\{f(\cdot; \theta)\}$ equivale a determinar cual es el verdadero valor $\theta_0$ de $\theta$. En este contexto $\theta$ se llama 'el parámetro', y el conjunto de valores posibles de $\theta$, denotado

por $\Theta$ es el espacio de parámetros. Consideraciones similares se aplican en el caso en que el vector de datos $\mathbf{Y}$ sea discreto. En este caso, el problema de estimación surge cuando lo que se sabe es que la función de probabilidad de $\mathbf{Y}$ pertenece a una familia $\{f(\cdot;\theta)\}$, donde

$$P[Y \in A] = \sum_{\mathbf{y} \in A} f(\mathbf{y};\theta) \tag{2.3}$$

El lado derecho de esta igualdad, o de (2.1), se denota mediante $P_\theta[A]$, y es la distribución de $\mathbf{Y}$ evaluada en $A$. El soporte de una densidad o función de probabilidad $f(\mathbf{y};\theta)$ es el conjunto de todos los vectores $\mathbf{y}$ en los que $f(\mathbf{y};\theta)$ es positiva:

$$\mathcal{Y}_\theta = \{\mathbf{y}\colon f(\mathbf{y};\theta) > 0\}.$$

En el desarrollo subsecuente, *se supondrá que todos los miembros de la familia* $\{f(\mathbf{y};\theta)\}$ *tienen el mismo soporte, esto es, que* $\mathcal{Y}_\theta$ *no dependa de* $\theta$.

## 1.3 El Problema

Suponga que se desea estimar la función $\psi(\theta) = \theta$. En este trabajo se aborda el problema de *demostrar la consistencia de los estimadores de verosimilitud máxima de* $\theta$; este método de estimación es ampliamente conocido y utilizado en las aplicaciones (Murphy y Topel 1985, Hardin 2002), y su consistencia es un resultado clásico ampliamente conocido en la literatrura (Dudewicz y Mishra 1988, Rao, 2002, Severini 2001), así que lo primero que viene a la mente es la razón para demostrar un resultado ya establecido. La razón es la siguiente:

> El resultado clásico sobre la consistencia de los estimadores de verosimilitud máxima supone que la familia de densidades $\{f(\mathbf{y};\theta)\}$ satisface ciertas condiciones de regularidad, las cuales, entre otros aspectos, involucran la existencia de derivadas hasta de tercer orden respecto a $\theta$ de las densidades $f(\mathbf{y};\theta)$ (Serfling 1988, Greene 2003, Grifiths *et. al.* 1997).

Las condiciones de regularidad fallan en modelos simples de traslación, en los que las densidades son de la forma $f(\mathbf{y};\theta) = g(\mathbf{y} - \theta)$ y $g$ tiene 'picos',

esto es, $g$ no es derivable en todo su dominio, y por lo tanto los resultados clásicos no se aplican. Por esta razón, es importante, e interesante desde el punto de vista estadístico, demostrar la consistencia de los estimadores de verosimilitud máxima bajo supuestos más débiles que las condiciones de regularidad usuales; este es el problema que se aborda en este trabajo, y presenta retos analíticos interesantes. Las condiciones que se utilizan en el desarrollo subsecuente involucran continuidad de $f(\mathbf{y};\theta)$ respecto a $\theta$ en el sentido de Lipschitz y, consecuentemente, son más débiles que los supuestos clásicos. Por lo tanto, este trabajo es una generalización de los resultados de consistencia conocidos.

## 1.4 Los Instrumentos

Como ya se ha mencionado, estudiar la consistencia de los estimadores de verosimilitud máxima sin suponer existencia de derivadas respecto a $\theta$ de $f(\mathbf{y};\theta)$, es un problema interesante, tanto desde el punto de vista estadístico como analítico. Las herramientas estadísticas que se utilizarán para analizar el problema son, esencialmente, dos: La ley de los grandes números, y la desigualdad de Jensen, la cual relaciona las ideas de valor esperado y de función cóncava. El instrumento analítico esencial es el Teorema de Heine-Borel, el cual se refiere a conjuntos compactos. Un conjunto contendio en $\mathbb{R}^k$ es compacto si es cerrado y acotado, y el toerema de Heine-Borel establece que al cubrir un conjunto compacto mediante una colección de conjuntos abiertos, existe una subcolección finita que tambien cubre al conjunto (Dugundji 1968, Munkres 1989, Rudin 1968).

## 1.5 La Organización

La presentación del material subsecuente ha sido organizada de la siguiente manera: En el Capítulo 2 se establece una versión de la ley fuerte de los grandes números que desempeñará un papel central en el desarrollo del trabajo. Esta formulación no supone que la esperanza de las variables aleatorias sea finita, sino que solamente se impone la condición de que la esperanza ex-

ista. Posteriormente se analiza la desigualdad de Jensen, prestando especial atención al caso de funciones estrictamente cóncavas. A continuación, se formula de manera precisa el método de verosimilitud máxima y se demuestra la consistencia de los estimadores que el método genera en el caso de que el espacio de parámetros es finito. Este resultado, aunque simple, ilustra las principales ideas que sustentan este trabajo.

En el Capítulo 3 se establecen los supuestos de compacidad y continuidad bajo las cuales se establece y se demustra la consistencia de los estimadores de verosimilitud máxima en el Teorema 3.3.1, el cual es la principal contribución de este trabajo. En el Capítulo 4 se proporciona el argumento clásico para establecer la consistencia de los estimadores de verosimilitud máxima. El proposito aquí, es ilustrar la complejidad analítica del enfoque clásico, y enfatizar en que puntos del argumento se utilizan los diversos supuesto de regularidad usuales. La demostración, sin embargo, es en cierto aspecto más simple que la que se encuentra disponible en la literatura, y está basada en la intoroducción de dos funciones cuadráticas cuyas raíces, en contraste con las soluciones de la ecuación de verosimilitud, pueden encontrarse fácilmente. El argumento en esta parte es una modificación del presentado en Serfflig (1988).

# Capítulo 2

# Consistencia

## 2.1 Introducción

Este capítulo trata sobre la idea central en este trabajo, a saber, la noción de consistencia de una sucesión de estimadores. Como punto de partida, sea $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ una sucesión de vectores aleatorios independientes con una distribución común, la cual se supone que tiene densidad $f(\mathbf{y}; \theta)$ respecto a una medida fija. En la aplicaciones, con frecuencia dicha medida es la de Lebesgue, en el caso continuo, o una medida de conteo, en el caso discreto. Por otro lado, $\theta$ es un parámetro desconocido, el cual pertenece a un conjunto $\Theta$. De esta manera, el observador no conoce exactamente la densidad de la distribución de los vectores $\mathbf{Y}_i$, 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 $\mathbf{Y}_i$ tiene densidad $f(\mathbf{y}; \theta_0)$; sin embargo, $\theta_0$ no es conocido por el observador, y su objetivo es utilizar los datos observados para estimar el valor de $\theta_0$ o, más generalmente, de una función $g(\theta_0)$. Un estimador de $g(\theta_0)$ basado en

$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ es una función

$$\hat{g}_n \equiv \hat{g}_n(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$$

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 2.1.1** Sea $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3 \ldots$ una sucesión de vectores aleatorios independientes e idénticamente distribuidos (iid), y suponga que la distribución común tiene densidad $f(\mathbf{y}; \theta)$ donde $\theta \in \Theta$. Denote mediante $\theta_0$ al verdadero valor del parámetro y sea $P_{\theta_0}$ la distribución correspondiente a $f(\mathbf{y}; \theta_0)$. Dada una función $g(\theta)$, una sucesión $\{\hat{g}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)\}$ es consistente si

$$P_{\theta_0}\left[\lim_{n\to\infty} \hat{g}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) = g(\theta_0)\right] = 1.$$

Esta noción se denomina consistencia fuerte en la literatura, para distinguirla de otra noción relacionada, llamada consistencia débil o consitencia en probabilidad, la cual requeire que, para cada $\varepsilon > 0$,

$$\lim_{n\to\infty} P_{\theta_0}\left[|\hat{g}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) = g(\theta_0)| > \varepsilon\right] = 0;$$

Para detalles sobre estas nociones vea, por ejemplo, Dudewicz y Mishra (1988), Mood *et. al.* (1987). Puede demostrarse que si una sucesión $\{\hat{g}_n\}$ es consistente en el sentido de la Definición 2.1.1, 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. Como ya se ha mencionado, el objetivo de este trabajo es establecer la consistencia de los estimadores de $\theta_0$ obtenidos mediante el método de verosimilitud máxima, y obtener este resultado bajo condiciones sobre la familia de

densidades $\{f(\mathbf{y};\theta)\}$ que son menos restrictivas que los requerimientos de regularidad usuales. En este capítulo se establecen los instrumentos que se utilizarán para obtener el principal resultado en esta dirección.

La exposición ha sido organizada de la siguiente manera: En la Sección 2 se presenta la ley de los grandes números, un importante resultado clásico que establece que los promedios muestrales de variables iid forma una sucesión consistente de estimadores de la media poblacional. En la Sección 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 $\log E[X] > E[\log(X)]$ se satisface cuando $X$ no es constante. El capítulo concluye en la Sección 4 introduciendo el método de verosimilitud máxima para construir estimadores del verdadero valor del parámetro $\theta_0$, y estableciendo la consistencia de dichos estimadores en modelos cuyo espacio de parámetros es finito. Aunque el resultado que se obtiene en este capítulo es simple, las ideas utilizadas para establecerlo serán útiles en la demostración del resultado general analizado más adelante.

## 2.2 Ley de los Grandes Números

El análisis de la consistencia de los estimadores de verosimilitud máxima depende fuertemente de la ley de los grandes números, resultado que se establece a continuación.

**Teorema 2.2.1** 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 $\mu = E[X_i]$.

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

$$P[\Omega^*] = 1 \quad \text{y} \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} = \mu, \quad \omega \in \Omega^*. \tag{2.1}$$

Una demostración de este resultado puede encontrarse, por ejemplo, en Ash (1987) o en Casella y Beger (2001) Usualmente, la conclusión (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 2.2.1, a saber, la condición de que las variables aleatorias $X_i$ tengan esperanza finita. Considere una variable aleatoria $X$ y defina

$$X^+ = \max\{X, 0\}, \quad \text{y} \quad X^- = \max\{-X, 0\}; \tag{2.2},$$

de modo que

$$X = X^+ - X^-; \tag{2.3}$$

$X^+$ 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 fórmulas 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

$$E[X^+] < \infty \quad \text{o} \quad E[X^-] < \infty \tag{2.4}$$

y en este caso, por definición

$$E[X] = E[X^+] - E[X^-]; \tag{2.5}$$

vea, por ejemplo Ash (1987), o DUdley (2002). Note que la condición (2.4) es necesaria para evitar que la fórmula para $E[X]$ en (2.5) resulte en '$\infty - \infty$', la cual no tiene sentido. Cuando $E[X^+] = \infty$ y $E[X^-] < \infty$, de acuerdo a la convención usual $\infty - a = \infty$ para todo $a \in \mathbb{R}$, la fórmula (2.5) arroja que $E[X] = \infty$. La única forma en que $E[X]$ sea finita, es que

$$E[X^+] < \infty \quad \text{y} \quad E[X^-] < \infty \tag{2.6}$$

pues en estas condiciones $E[X]$ en (2.5) es la diferencia de dos números reales (i.e., finitos). El supuesto de que las variables aleatorias $X_i$ en el Teorema 2.2.1 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 2.2.2** 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 una evento $\Omega^* \subset \Omega$ tal que

$$P[\Omega^*] = 1 \quad \text{y} \quad \lim_{n\to\infty} \frac{\sum_{i=1}^n X_i(\omega)}{n} = \mu, \quad \omega \in \Omega^*. \tag{2.7}$$

Note que la única diferencia entre las condiciones de los Teorema 2.2.1 y 2.2.2 es que en este último teorema la esperanza común de las variables $X_i$ puede ser $\infty$ o $-\infty$. Si, por ejemplo, $E[X_i] = \infty$, entonces el Teorema 2.2.2 asegura que $\lim_{n\to\infty}(X_1 + \cdots X_j n)/n = \infty$ 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.** 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 2.2.1. Con esto en mente, suponga que $E[X_i] = -\infty$, esto es, que

$$\mu^+ = E[X_i^+] < \infty \quad \text{y} \quad E[X_i^-] = \infty. \tag{2.8}$$

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

$$\tilde{X}_{i,N} := \min\{X_i^-, N\}. \tag{2.9}$$

como $X_i^- \geq 0$ (vea (2.2)) se tiene que

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

y entonces

$$\tilde{\mu}_N := E[\tilde{X}_{i,N}] \leq N. \tag{2.10}$$

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

$$\tilde{X}_{i,N} \leq \tilde{X}_{i,N+1} \leq X_i^-, \quad N = 1, 2, 3, \ldots \quad \text{y} \quad \lim_{N\to\infty} \tilde{X}_{i,N} = X_i^- \tag{2.11}$$

y a partir del teorema de convergencia monótona se concluye que

$$\lim_{N\to\infty} \tilde{\mu}_N = \lim_{N\to\infty} E[\tilde{X}_{i,N}] = E[X_i^-] = \infty. \tag{2.12}$$

Defina ahora

$$X_{i,N} = X_i^+ - \tilde{X}_{i,N}. \tag{2.13}$$

Para cada $N$ fijo, las vartiables 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 (2.8) y (2.10). Luego, aplicando el Teorema 2.2.1, se desprende que existe un evento $\Omega_N^*$ con

$$P[\Omega_N^*] = 1 \tag{2.14}$$

tal que

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n X_{i,N}(\omega)}{n} = \mu^+ - \tilde{\mu}_N, \quad \omega \in \Omega_N^*. \tag{2.15}$$

Por otro lado, usando que $X_i^- \geq \tilde{X}_{i,N}$ (vea (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 (2.13) y (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 (2.15) esta relación conduce a

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} X_i(\omega)}{n} \leq \mu^+ - \tilde{\mu}_N, \quad \omega \in \Omega_N^*. \tag{2.16}$$

Para concluir, defina $\Omega^* = \bigcap_{N=1}^{\infty} \Omega_N^*$. En este caso, (2.14) implica que

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

mientras que combinando la inclusión $\Omega^* \subset \Omega_N^*$ con (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. □

## 2.3 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 $t \in [0,1]$,

$$g(tx_1 + (1-t)x_2) \geq tg(x_1) + (1-t)g(x_2). \tag{3.1}$$

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 (3.1) 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 totalmenta arbitraria, y sólo se supone que sus valores pertenecen al intervalo $I$ en el cual la función $g$ es cóncava.

**Teorema 2.3.1** Considere un intervalo $I$ de $\mathbb{R}$ y sea $g: I \to \mathbb{R}$ una función cóncava en $I$. Supoga 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

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

Para una demostración de este resultado vea, por ejemplo, Ash (1987) o Dudewicz y Mishra (1989) . En el análisis de la consistencia de los estimadores de verosimilitud máxima se utilizará un caso especial del Teorema 2.3.1, 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 (3.1), esto es, $g(tx_1 + (1-t)x_2) > tg(x_1) + (1-t)g(x_2)$. 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 fucnión $g(x) = \log(x)$ es estrictamente cóncava en $I = (0, \infty)$, pues $g''(x) = -1/x^2 < 0$ para todo $x > 0$.

**Teorema 2.3.2** Suponga que $g$ es una función estrictamente cóncava en un intervalo $I$. Sea $X$ una variable aleatoria para la cual $P[X \in I] = 1$ y cuya esperanza $\mu$ es finita. Si $X$ no es constante con probabilidad 1, esto es, si $P[X = \mu] < 1$, entonces

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

Una demostración de este resultado puede verse, por ejemplo, en Rudin (1968), Dudley (2001) o en Ash (1987). El siguiente caso especial del Teorema 2.3.2, 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 2.3.1** Sea $X$ tal que $P[X > 0] = 1$. Suponga que $E[X] = 1$ y que $P[X = 1] < 1$. En este caso,

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

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

**Ejemplo 2.3.1** [Cociente de densidades.] Sea $\mathbf{Y}$ un vector aleatorio con densidad $f_0(\mathbf{y})$ y denote mediante $\mathcal{Y}$ al soporte de $f_0(\mathbf{y})$, esto es,

$$\mathcal{Y} = \{\mathbf{y} \colon f_0(\mathbf{y}) > 0\}.$$

En este caso, para cada función $H(\mathbf{Y})$ para la cual $E[H(\mathbf{Y})]$ está definida,

$$E[H(\mathbf{Y})] = \int_{\mathcal{Y}} H(\mathbf{y}) f_0(\mathbf{y})\, d\mathbf{y}. \tag{3.2}$$

Considere ahora otra densidad $f_1(\mathbf{y})$ cuyo soporte es $\mathcal{Y}$, de manera que $f_1(\mathbf{y}) > 0$ para $\mathbf{y} \in \mathcal{Y}$ y

$$\int_{\mathcal{Y}} f_1(\mathbf{y})\, d\mathbf{y} = 1,$$

y suponga que las densidades $f_1$ y $f_0$ son distintas, en el sentido de que $\int_A f_1(\mathbf{y})\, d\mathbf{y} \neq \int_A f_0(\mathbf{y})\, d\mathbf{y}$ son diferentes para algún intervalo $A$. Considere ahora el cociente $X$ de las densidades $f_1$ y $f_0$ evaluadas en la variable aleatoria $Y$, esto es,

$$X = \frac{f_1(Y)}{f_0(Y)} = H(\mathbf{Y}).$$

En este caso,

(i) $P[X > 0] = 1$, pues $Y$ toma valores en $\mathcal{Y}$ con probabilidad 1 y tanto $f_1$ como $f_0$ son positivas en $\mathcal{Y}$.

(ii) Aplicando (3.2) se obtiene que

$$\begin{aligned} E[X] &= E[H(Y)] \\ &= \int_{\mathcal{Y}} H(y) f_0(\mathbf{y})\, d\mathbf{y} \\ &= \int_{\mathcal{Y}} \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} f_0(\mathbf{y})\, d\mathbf{y} \\ &= \int_{\mathcal{Y}} f_1(\mathbf{y})\, d\mathbf{y} \\ &= 1. \end{aligned}$$

(iii) Como las densidades $f_0$ y $f_1$ son distintas, se tiene que $P[X = 1] < 1$. En efecto, si $P[X = 1] = 1$, entonces $P[X - 1 = 0] = 1$, de menara que, para cada intervalo $A$,

$$P[(X - 1) I_A(\mathbf{Y}) = 0] = 1,$$

de donde se desprende que

$$0 = E[(X - 1) I_A(Y)] = E\left[\left(\frac{f_1(\mathbf{Y})}{f_0(\mathbf{Y})} - 1\right) I_A(\mathbf{Y})\right]$$

y, via (3.2) se concluye que

$$\begin{aligned} 0 &= \int_{\mathcal{Y}} \left(\frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} - 1\right) I_A(\mathbf{y}) f_0(\mathbf{y})\, d\mathbf{y} \\ &= \int_{\mathcal{Y}} \left(f_1(\mathbf{y}) - f_0(\mathbf{y})\right) I_A(\mathbf{y})\, d\mathbf{y} \\ &= \int_A (f_1(\mathbf{y}) - f_0(\mathbf{y})\, d\mathbf{y}, \end{aligned}$$

y entonces $\int_A f_1(\mathbf{y})\,d\mathbf{y} = \int_A f_0(\mathbf{y})\,d\mathbf{y}$; como este argumento es válido para todo intervalo $A$, se tiene una contradicción con el supuesto de que $f_0$ y $f_1$ son densidades distintas. Por lo tanto $P[X = 1] < 1$, como se afirmó.

Las propiedades (i)–(iii) muestran que $X$ satisface las condiciones del Corolario 2.3.1, de tal manera que

$$0 > E[\log(X)] = E\left[\log\left(\frac{f_1(Y)}{f_0(Y)}\right)\right] = \int_{\mathcal{Y}} \log\left(\frac{f_1(\mathbf{y})}{f_0(\mathbf{y})}\right) f_0(\mathbf{y})\,d\mathbf{y}.$$

En resumen: Si $\mathbf{Y}$ tiene densidad $f_0(\mathbf{y})$ y $f_1(\mathbf{y})$ es cualquier densidad distinta con el mismo soporte que $f_0(\mathbf{y})$, entonces $X = \log(f_1(Y)/f_0(Y))$ tiene esperanza negativa. □

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

**Ejemplo 2.3.2** 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

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

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

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

□

Después de los resultados técnicos preliminares en esta y la sección precedente, a continuación se aborda el tema central de este trabajo.

## 2.4 El Método de Verosimilitud Máxima: Motivación

Suponga que $\mathbf{Y}$ es un vector aleatorio cuya densidad pertence a la familia $\{f(\mathbf{y};\theta):\theta \in \Theta\}$, y denote por $\theta_0$ al verdadero valor el parámetro, el cual es desconocido para el observador, de manera que la densidad del vector $\mathbf{Y}$ es $f(\mathbf{y};\theta_0)$. Suponga, además que el soporte de las densidades $f(\mathbf{y};\theta)$ no depende de $\theta$, y que la siguiente condición es válida.

**Hipótesis 2.4.1** [Identificabilidad.] *Densidades correspondientes a parámetros distintos son diferentes, i.e., si $\theta \neq \theta_1$, entonces existe un conjunto $A$ tal que*

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

Como es usual, se utilizará la expresión $E_\theta[H(Y)]$ para indicar la esperanza de la variable aleatoria $Y$ bajo el supuesto de que la densidad de $\mathbf{Y}$ es $f(\mathbf{y};\theta)$, de manera que $E_{\theta_0}[H(\mathbf{Y})]$ denota la esperanza de $H(\mathbf{Y})$ cuando el valor del parámetro es $\theta_0$, el verdadero valor. Para obtener información sobre $\theta_0$, el cual se desconoce, el observador toma una muestra $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \ldots, \mathbf{Y}_n$ y trata de obtener una buena estimación a partir de los datos observados. Lo que el observador sabe es que esta muestra se obtuvo a partir de una densidad $f(\mathbf{y};\theta)$ donde $\theta \in \Theta$, pero no sabe que el valor real de $\theta$ es $\theta_0$. Así, él conoce que la densidad de $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ es

$$f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n;\theta) = \prod_{i=1}^{n} f(\mathbf{y}_i;\theta), \quad \theta \in \Theta. \tag{4.1}$$

Sin embargo, el observador no sabe el valor preciso de $\theta$, es decir, no sabe que la verdadera densidad de $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ es

$$f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n;\theta_0) = \prod_{i=1}^{n} f(\mathbf{y}_i;\theta_0). \tag{4.2}$$

El método de verosimilitud máxima para construir un estimador de $\theta$ está basado en el siguiente resultado.

**Teorema 2.4.1** (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 $P_{\theta_0}$,

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

(iii) Existe un evento $\Omega^*$ tal que $P[\Omega^*] = 1$ para el cual la siguiente afirmación es cierta: Para cada $\omega \in \Omega^*$, existe un entero $N(\omega;\theta)$ tal que

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

o, equivalentemente,

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

**Demostración.** (i) Esta parte se desprende del Ejemplo 2.3.1 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, via (4.1) y (4.2), se obtiene que

$$\log\left(\frac{f_n(\mathbf{Y}_1,\ldots,\mathbf{Y}_n;\theta)}{f_n(\mathbf{Y}_1,\ldots,\mathbf{Y}_n;\theta_0)}\right) = \log\left(\frac{\prod_{i=1}^n f(\mathbf{Y}_i;\theta)}{\prod_{i=1}^n f(\mathbf{Y}_i;\theta_0)}\right) = \sum_{i=1}^n \log\left(\frac{f(\mathbf{Y}_i;\theta)}{f(\mathbf{Y}_i;\theta_0)}\right). \tag{4.3}$$

Por otro lado, las variabes 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 $P_{\theta_0}$, 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 2.2.2 implica que existe un evento $\Omega^*$ con $P_{\theta_0}[\Omega^*] = 1$, tal que

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

Combinando esta relación con (4.3) se desprende que

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

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

(iii) Sea $\Omega^*$ 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 $\omega \in \Omega^*$, existe un entero $N(\omega, \theta)$ tal que

$$\frac{1}{n} \log \left( \frac{f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta)}{f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0)} \right) < 0, \quad n > N(\omega, \theta),$$

lo cual equivale a

$$\begin{aligned} &\log \left( f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta) \right) \\ &\qquad < \log \left( f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0) \right), \quad n > N(\omega, \theta), \end{aligned}$$

o bien

$$\begin{aligned} &f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta) \\ &\qquad < f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0), \quad n > N(\omega, \theta) \end{aligned},$$

concluyendo la demostración. □

Dados los datos observados $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, defina

$$L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = f_n(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n; \theta), \quad \theta \in \Theta, \tag{4.4}$$

la cual se denomina función de verosimilitud. Si los vectores $\mathbf{Y}_i$ son discretos, el valor de $L_n(\theta; \mathbf{y}_1, \ldots, \mathbf{y}_n)$ es la probabilidad de observar el evento $[\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2, \ldots, \mathbf{Y}_n = \mathbf{y}_n]$ cuando $\theta$ es el valor del parámetro. El hecho fundamental que se establece en la parte (iii) del Teorema anterior, es que, para $n$ suficientemente grande, la función de verosimilitud $L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ no se maximiza en $\theta \neq \theta_0$, por lo cual es razonable buscar 'buenos' estimadores del valor desconocido $\theta_0$ entre los maximizadores de la función de verosimilitud.

**Definición 2.4.1** [Método de Verosimilitud Máxima.] Sea $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ una muestra de una distribución con densidad $f(\mathbf{y}; \theta)$, donde $\theta \in \Theta$. Un estimador $\hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in \Theta$ se denomina estimador de verosimilitud máxima si

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

El siguiente resultado muestra que si el espacio de parámetros es finito, entonces una sucesión $\{\hat{\theta}_n\}$ de estimadores de verosimilitud máxima es consistente. Note que cuando el espacio de parámetros $\Theta$ es finito, entonces $L(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ siempre tiene un maximizador como función de $\theta \in \Theta$, de modo que $\hat{\theta}_n$ está siempre definido.

**Teorema 2.4.2** Sea $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ una muestra de una distribución con densidad $f(\mathbf{y}; \theta)$, donde $\theta \in \Theta$, y el espacio de parámetros $\Theta$ es finito. Denote por $\theta_0 \in \Theta$ al verdadero valor del parámetro y suponga que la Hipótesis 2.4.1 es válida. En este caso, existe un evento $\Omega^*$ tal que

(i) $P_{\theta_0}[\Omega^*] = 1$,

(ii) Para cada $\omega \in \Omega^*$, existe un entero $N(\omega)$ tal que

$$\hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots \mathbf{Y}_n(\omega)) = \theta_0, \quad n > N(\omega).$$

Este resultado muestra que, con probabilidad 1, cuando el tamaño de la muestra es suficientemente grande, el estimador de verosimilitud máxima coincide con el verdadero valor del parámetro y, por lo tanto, la sucesión $\{\hat{\theta}_n\}$ de estimadores de verosimilitud máxima converge con probabilidad 1 al verdadero valor del parámetro $\theta_0$, ie.,

$$P_{\theta_0}\left[\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right] = 1$$

**Demostración.** Escriba

$$\Theta = \{\theta_0, \theta_1, \ldots, \theta_r\}, \tag{4.5}$$

donde $\theta_0$ es el verdadero valor del parámetro. Para cada $i = 1, 2, \ldots, r$, por el Teorema 2.4.1(iii) existe un evento $\Omega_i^*$ tal que

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

(ii) Para cada $\omega \in \Omega_i^*$ existe un entero $N(\omega, \theta_i)$ tal que

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

lo cual equivale a

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

Esta desigualdad muestra que

$$\text{Para } \omega \in \Omega_i^*, \quad \hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) \neq \theta_i \quad \text{si } n > N(\omega, \theta_i). \tag{4.6}$$

Ahora defina

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

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

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

y observe que $N(\omega)$ es finito, pues el espacio de parámetros lo es. Seleccione ahora $\omega \in \Omega^*$ y tome $n > N(\omega)$. En este caso $\omega \in \Omega_i^*$ y $n > N(\omega, \theta_i)$ para cada $i = 1, 2, \ldots, r$, y (4.6) implica que $\hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) \neq \theta_i$. En resumen:

$$\text{Si } \omega \in \Omega^* \text{ y } n > N(\omega), \quad \text{entonces } \hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) \neq \theta_i, \quad i = 1, 2, \ldots, r$$

Como $\hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) \in \Theta$, combinando este enunciado con (4.5) se desprende que

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

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

El objetivo de este trabajo es extender la conclusión del Teorema 2.4.2 a un marco de trabajo más general, en el cual el espacio de parámetros no es finito. Esta es la tarea que se desarrolla en los siguientes capítulos

# Capítulo 3

# Estimación en el Caso de Espacio de Parámetros Compacto

## 3.1 Introducción

En este capítulo se establece el principal resultado de este trabajo, a saber, la consistencia de los estimadores de verosimilitud máxima bajo condiciones de continuidad respecto al parámetro de la función de densidad. Estas condiciones no involucran diferenciabilidad, y por lo tanto son más déblies que las condiciones de regularidad usuales. Por otro lado, es un hecho conocido que una función puede no tener un maximizador si no se imponen condiciones sobre su dominio. Para una función $L$ definida en un subconjunto de $\mathbb{R}^n$, una condición necesaria para que $L$ tenga maximizador es que su dominio sea compacto, es decir, que sea cerrado y acotado, y se supondrá la compacidad del espacio de parámetros. El resultado principal de este trabajo se formula en el Teorema 3.3.1, y la estrategia para demostrarlo se basa en tres resultados fundamentales:

(i) La ley de los grandes números, la cual establece que el promedio muestral

de variables aleatorias independientes con distribución común converge con probabilidad 1 hacia la media poblacional; vea la formulación presisa en el Teorema 2.2.2 del Capítulo 2.

(ii) La desigualdad de Jensen, la cual relaciona las ideas de esperanza y de función cóncava; particularmente, se usará el resultado sobre el valor esperado del logaritmo de un cociente de densidades en Ejemplo 2.3.1.

(iii) El teorema de Heine-Borel, también llamado el teorema de subcubiertas finitas para conjuntos compactos. Este resultado es una propiedad fundamental del sistema de números reales y será formalmente establecido en este capítulo.

La organización del material del capítulo es la sigiente: En la Sección 2 se introducen los supuestos de continuidad y compacidad bajo los cuales se analizará la consistencia de los estimadores de verosimilitud máxima; además se incluyen dos ejemplos para ilustrar la verificación de las hipótesis en dos casos comunes en las aplicaciones. En uno de los ejemplos, las condiciones usuales de diferenciabilidad no se satisfacen. En la Sección 3 se establece el resultado de consistencia en el Teorema 3.3.1, mientras que la demostración se presenta en la Sección 4.

## 3.2 Supuestos de Continuidad–Compacidad

Para determina el estimador $\hat{\theta}_n$ de $\theta_0$ es necesario maximizar la función de verosimilitud $L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ como función de $\theta \in \Theta$. Para garantizar que esta función tenga realmente un maximizador, de manera que $\hat{\theta}_n$ esté bien definido, es necesario imponer condiciones, tanto sobre el dominio de la función—en este caso $\Theta$—como sobre la continuidad de la función. Recuerde que un subconjunto no vacío $\Theta \subset \mathbb{R}^k$ es cerrado si para cada sucesión $\{\theta_n\} \subset \Theta$ se tiene que si $\lim_{n\to\infty} \theta_n = \theta$ entonces $\theta \in \Theta$. Además, $\Theta$ es acotado si existe una constante $B > 0$ tal que $\|\theta\| \leq B$ para todo $\theta \in \Theta$, mientras que $\Theta$ es compacto si es tanto cerrado como acotado. En términos generales, para garantizar que una función tenga un maximizador en su dominio es necesario suponer que éste es compacto, razón por la cual se impone la

siguiente condición.

**Hipótesis 3.2.1** *El espacio de parámetros* $\Theta$ *es un subconjunto compacto de* $\mathbb{R}^k$.

Este supuesto tiene una consecuencia que desempeña un importante papel en el análisis subsecuente. Para cada punto $\mathbf{x} \in \mathbb{R}^k$ y $\varepsilon > 0$, defina la bola con centro $\mathbf{x}$ y radio $\varepsilon > 0$ mediante

$$B(\mathbf{x}, \varepsilon) := \{\mathbf{y} \in \mathbb{R}^k : \|\mathbf{y} - \mathbf{x}\| < \varepsilon\}. \tag{2.1}$$

**Teorema 3.2.1** Sea $\Theta$ un subconjunto compacto de $\mathbb{R}^k$. Considere un subconjunto $F \subset \mathbb{R}^k$ y para cada $\mathbf{x} \in F$ sea $\varepsilon_{\mathbf{x}}$ un número positivo. Suponga que

$$\Theta \subset \bigcup_{\mathbf{x} \in F} B(\mathbf{x}, \varepsilon_{\mathbf{x}})$$

En este caso, existe un subconjunto finito $\{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_m\} \subset F$ tal que

$$\Theta \subset \bigcup_{i=1}^{m} B(\mathbf{x}_i, \varepsilon_{\mathbf{x}_i})$$

Este resultado, que es una de las propiedades más profundas del sistema de números reales, se conoce como teorema de Heine-Borel, o teorema de la subcubierta finita. Una demostración puede encontrarse en Ash (1987), Dugundji (1968), o Munkres (1989). Además de la compacidad de su dominio, para garantizar que una función tenga un maximizador es necesario suponer que es continua, propiedad que efectivamente se verifica cuando la función es diferenciable. Las condiciones clásicas sobre la función de verosimilitud suponen que ésta es diferenciable como función del parámetro $\theta$. En este trabajo se supondrá la siguiente condición sobre la dependencia de la función de verosimilitud respecto al parámetro $\theta$, la cual es más fuerte que el simple supuesto de continuidad, pero que es más débil que el requerimiento clásico de diferenciabilidad. En este punto es conveniente recordar

que la discusión en este trabajo supone que todas las densidades de la familia $\{f(\mathbf{y};\theta)\colon \theta \in \Theta\}$ tienen el mismo soporte $\mathcal{Y}$, esto es,

$$\mathcal{Y} := \{\mathbf{y}\colon f(\mathbf{y};\theta) > 0\}$$

es el mismo para cada $\theta \in \Theta$.

**Hipótesis 3.2.2** [Condiciones de Continuidad.] *Existe una función* $B\colon \mathcal{Y} \to [0,\infty)$ *tal que las siguientes condiciones (i) y (ii) son válidas:*

*(i) La función* $\theta \mapsto \log f(\mathbf{y};\theta)$ *es continua con modulo de continuidad* $B(\mathbf{y})$, *esto es,*

$$|\log f(\mathbf{y};\theta) - \log f(\mathbf{y};\theta_1)| \leq \|\theta - \theta_1\| B(\mathbf{y}), \quad \theta, \theta_1 \in \Theta, \quad \mathbf{y} \in \mathcal{Y}. \tag{2.2}$$

*Más aún,*

*(ii) La variable aleatoria* $B(Y)$ *tiene esperanza finita sin importar cual sea el verdadero valor del parámetro, i.e.,*

$$\mu_B(\theta) := E_\theta[B(Y)] < \infty, \quad \theta \in \Theta. \tag{2.3}$$

Las condición (2.2) establece que, como función de $\theta$, $f(\mathbf{y};\theta)$ es una función continua en el sentido de Lipschitz, con constante de Lipschitz $B(\mathbf{y})$. Por otro lado, como ya se ha mencionado, la estrategia que se seguirá para analizar la consistencia de los estimadores de verosimilitud máxima gira alrededor de la ley de los grandes números, y el uso de este último resultado requiere que la condición (2.3) sea válida. A continuación se discute la validez de la Hipóteisis 3.2.2 en algunos modelos comunes en las aplicaciones.

**Ejemplo 3.2.1** Considere la familia de densidades dada por

$$f(\mathbf{y};\xi) = a(\xi)^{-1} h(\mathbf{y}) e^{\xi' T(\mathbf{y})}, \quad \mathbf{y} \in \mathbb{R}^s, \quad \xi \in \Xi$$

donde $T(\mathbf{y}) = [T_1(\mathbf{y}), T_2(\mathbf{y}), \ldots, T_k(\mathbf{y})]'$, $\xi \in \mathbb{R}^k$,

$$a(\xi) = \int_{\mathbb{R}^k} h(\mathbf{y}) e^{\xi' T(\mathbf{y})}\, d\mathbf{y},$$

mientras que $\Xi$ consiste de todos los vectores en $\mathbb{R}^k$ para los cuales $a(\xi)$ es finita. En este caso $\Xi$ es un conjunto convexo y, sin pérdida de generalidad, se supondrá que $\Xi$ no tiene interior vacío. El conjunto $\{f(\mathbf{y};\xi)\colon \xi \in \Xi\}$ es una familia exponencial $k$-parametral para la cual las siguientes propiedades son válidas en cada punto interior $\xi_0$ de $\Xi$:

(i) Cada componente $T_i(Y)$ de $T(Y)$ tiene esperanza finita respecto a $f(\cdot;\xi_0)$, esto es,

$$E_{\xi_0}[|T_i(Y)|] < \infty, \quad i = 1, 2, \ldots, k. \tag{2.4}$$

(ii) La función $a(\cdot)$ es derivable en $\xi_0$ y

$$\partial_{\xi_r} a(\xi_0) = a(\xi_0) E_{\xi_0}[T_r(Y)], \quad r = 1, 2, \ldots, k.$$

Más aún,

(iii) Las derivadas parciales de $a(\xi)$ son continuas en todo el interior de $\Xi$.

Una demostración de estos resultados puede encontrarse en Lehmman (2001). Note que en este caso $\mathcal{Y}$, el conjunto soporte de $f(\mathbf{y};\xi)$ es

$$\mathcal{Y} = \{\mathbf{y}\colon h(\mathbf{y}) > 0\}.$$

Ahora, sea $\Theta$ un conjunto compacto contenido en el interior de $\Xi$. En este caso, se verificará a continuación que la familia $\{f(\mathbf{y};\xi)\colon \xi \in \Theta\}$ satisface los requerimientos en la Hipótesis 3.2.2. Con este fin note que para $\mathbf{y} \in \mathcal{Y}$

$$\begin{aligned} \log f(\mathbf{y};\xi) &= \log[a(\xi)^{-1} h(\mathbf{y}) e^{\xi' T(\mathbf{y})}] \\ &= \xi' T(\mathbf{y}) + \log(h(\mathbf{y})) - \log(a(\xi)), \end{aligned}$$

y entonces

$$\log f(\mathbf{y};\xi) - \log f(\mathbf{y};\xi_1) = (\xi - \xi_1)' T(\mathbf{y}) + \log[a(\xi)] - \log[a(\xi_1)], \quad \xi, \xi_1 \in \Xi, \quad \mathbf{y} \in \mathcal{Y}. \tag{2.5}$$

Note ahora que, por la desigualdad de Cauchy-Schwarz,

$$|(\xi - \xi_1)' T(\mathbf{y})| \leq \|\xi - \xi_1\| \, \|T(\mathbf{y})\|$$

donde $\|\mathbf{w}\|$ denota la norma Euclideana del vector $\mathbf{w}$, de manera que

$$\|T(\mathbf{y})\| = \sqrt{T_1(\mathbf{y})^2 + \cdots + T_k(\mathbf{y})^2} \leq |T_1(\mathbf{y})| + \cdots + |T_k(\mathbf{y})|,$$

y por lo tanto

$$|(\xi - \xi_1)'T(\mathbf{y})| \leq \|\xi - \xi_1\| \, \|T(\mathbf{y}) = \|\xi - \xi_1\| \, [|T_1(\mathbf{y})| + \cdots + |T_k(\mathbf{y})|] \quad (2.6)$$

Ahora se establecerá una cota para la diferencia de logaritmos en (2.5). Sea $\tilde{\Theta}$ la unión de todos los segmentos que unen puntos de $\Theta$, esto es, $\tilde{\Theta} = \{t\theta + (1-t)\theta_1 \,|\, \theta, \theta_1 \in \Theta, \; t \in [0,1]\}$. Como $\Theta$ es un conjunto compacto contenido en el interior de $\Xi$, el cual es un conjunto convexo, se desprende $\tilde{\Theta}$ es compacto y está contenido en el interior de $\Xi$. Luego, las derivadas parciales de $a(\xi)$, las cuales están definidas y son continuas en el interior de $\Xi$, estan acotadas cuando $\xi \in \tilde{\Theta}$, i.e., existe $\tilde{M} > 0$ tal que $|\partial_{\xi_i} a(\xi)| \leq \tilde{M}$, $i = 1, 2, \ldots, k$, $\xi \in \tilde{\Theta}$. Por lo tanto,

$$\begin{aligned} \|Da(\xi)\| &= \sqrt{\partial_{\xi_1} a(\xi)^2 + \cdots + \partial_{\xi_k} a(\xi)^2} \\ &\leq \sqrt{\tilde{M}^2 + \cdots + \tilde{M}^2} = \sqrt{k}\tilde{M} =: M, \quad \xi \in \tilde{\Theta}. \end{aligned}$$

Seleccione ahora $\xi, \xi_1 \in \Theta$ . Por el teorema del valor medio, existe $t \in (0,1)$ tal que

$$a(\xi) - a(\xi_1) = Da(t\xi + (1-t)\xi_1)(\xi - \xi_1);$$

en este caso $t\xi + (1-t)\xi_1 \in \tilde{\Theta}$, y la anterior deisgualdad desplegada implica que

$$\|Da(t\xi + (1-t)\xi_1)\| \leq M,$$

y aplicando la desigualdad de Cauchy-Schwarz se concluye que

$$\begin{aligned} |a(\xi) - a(\xi_1)| &= |Da(t\xi + (1-t)\xi_1)(\xi - \xi_1)| \\ &\leq \|Da(t\xi + (1-t)\xi_1)\| \, \|\xi - \xi_1\|; \qquad (2.7) \\ &\leq M\|\|\xi - \xi_1\|, \quad \xi, \xi_1 \in \Theta. \end{aligned}$$

Por otro lado, debido a que $a(\xi)$ es continua y positiva para $\xi$ en el interior de $\Xi$, el cual contiene al conjunto compacto $\tilde{\Theta}$, existe una constante $b > 0$ tal que

$$a(\xi) > b, \quad \xi \in \Theta.$$

Usando que $\log'(x) = 1/x$, el teorema del valor medio implica que existe una constante $s \in (0,1)$ tal que

$$\log a(\xi) - \log a(\xi_1) = \frac{1}{sa(\xi) + (1-s)a(\xi_1)}[a(\xi) - a(\xi_1)];$$

cuando $\xi$ y $\xi_1$ pertenecen a $\Theta$ se tiene que $a(\xi) > b$ y $a(\xi_1) > b$ y por lo tanto la inclusión $s \in (0,1)$ implica que $sa(\xi) + (1-s)a(\xi_1) > b$. Combinando este hecho con la anterior relación desplegada se obtiene que

$$|\log a(\xi) - \log a(\xi_1)| \leq \frac{1}{b}|a(\xi) - a(\xi_1)|, \quad \xi, \xi_1 \in \Theta$$

y junto con (2.7) esto implica que

$$|\log a(\xi) - \log a(\xi_1)| \leq \frac{M}{b}\|\xi - \xi_1\|, \quad \xi, \xi_1 \in \Theta.$$

Combinando esta desigualdad con (2.5) y (2.6) se obtiene

$$|\log f(\mathbf{y};\xi) - \log f(\mathbf{y};\xi_1)| \leq \left[|T_1(\mathbf{y})| + \cdots + |T_k(\mathbf{y})| + \frac{M}{b}\right]\|\xi - \xi_1\|, \quad \xi, \xi_1 \in \Theta, \quad \mathbf{y} \in \mathcal{Y}.$$

Definiendo $B(\mathbf{y}) := |T_1(\mathbf{y}) + \cdots + T_k(\mathbf{y})] + \frac{M}{b}$ se tiene que la primera parte de la Hipótesis 3.2.2 ocurre (vea (2.2)), mientras que la desigualdad (2.4), válida cuando $\xi_0$ es un punto interior de $\Xi$, implica que

$$E_\theta\left[B(\mathbf{Y})\right] = E_\theta\left[|T_1(\mathbf{Y})|\right] + \cdots + E_\theta\left[|T_k(\mathbf{Y})|\right] + \frac{M}{b} < \infty, \quad \theta \in \Theta,$$

puesto que $\Theta$ está contenido en el interior de $\Xi$. Por lo tanto, también la segunda parte de la Hipótesisi 3.2.2 es válida. □

El siguiente ejemplo se refiere a una familia de densidades que no es una familia exponencial, y muestra que las condiciones en la Hipóteis 3.2.2 pueden ser satisfechas aún en el caso en que $\log f(\mathbf{y};\theta)$ no sea diferenciable respecto a $\theta$, como lo requieren las condiciones clásicas.

**Ejemplo 3.2.2** Sea $f(y;\xi)$ la densidad de Laplace con centro $\xi$, i.e.,

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

En este caso $\log f(y;\xi) = -|y-\xi| - \log 2$, y entonces

$$\log f(y;\xi) - \log f(y;\xi_1) = |y-\xi| - |y-\xi_1|;$$

usando la desigualdad del triángulo, la cual establece que

$$||a| - |b|| \leq |a-b|$$

para todos los números reales $a$ y $b$, se desprende que

$$||y-\xi| - |y-\xi_1|| \leq |(y-\xi) - (y-\xi_1)| = |\xi_1 - \xi|$$

y entonces

$$|\log f(y;\xi) - \log f(y;\xi_1)| \leq |\xi - \xi_1|;$$

definiendo $B(y) = 1$ para todo $y \in \mathbb{R}$, se desprende que la Hipótesis 3.2.2 se satisface. □

Después de establecer los instrumentos técnicos necesarios, se está en posibilidad de formular y demostrar la principal contribución de este trabajo

## 3.3 El Resultado Principal

En esta sección se formula el resultado de consistencia de los estimadores de verosimilitud máxima bajo los supuestos de continuidad–compacidad establecidos anteriormente. Como punto de partida note que, bajo la las Hipótesis 3.2.1 y 3.2.2 la función de verosimilitud $L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ depende continuamente de $\theta$, y por lo tanto tiene un maximizador, $\hat{\theta}_n = \hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ pues el espacio de parámetros es compacto. Esto significa que siempre es posible encontrar un estimador de verosimilitud máxima basado en $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$. La principal contribución de este trabajo se establece a continuación.

**Teorema 3.3.1** Sea $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \ldots$ una sucesión de vectores aleatorios independientes con una distrinución común cuya densidad pertenece a la familia $\{f(\mathbf{y};\theta): \theta \in \Theta\}$ y suponga que las Hipótesis 3.2.1 y 3.2.2 son válidas. Denote por $\theta_0$ al verdadero valor del parámetro y sea $\hat{\theta}_n$ un estimador de verosimilitud máxima basado en $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$. En estas circunstancias,

$$P_{\theta_0}\left[\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right] = 1. \tag{3.1}$$

La conclusión (3.1) es que los estimadores de verosimilitud máxima estiman consistentemente al verdadero valor del parámetro $\theta_0$; ésta, desde luego, es la conclusión clásica, pero la diferencia es que las condiciones bajo las cuales se obtiene, a saber, las Hipótesis 3.2.1 y 3.2.2 son más débiles que las condiciones de regularidad usuales (Greene 2003, Grifiths *et. al.* 1997). La demostración del Teorema 3.3.1 es algo técnica y se presenta en la última sección de este capítulo. El argumento utiliza el siguiente resultado que es una extensión del Teorema 2.4.1 del Capítulo 2

**Teorema 3.3.2** Suponga que las Hipótesis 3.2.1 y 3.2.2 son válidas y sea $\theta^* \in \Theta$ un parámetro arbitrario, con

$$\theta^* \neq \theta_0$$

En este caso, existe un evento $\Omega^*$ con $P[\Omega^*] = 1$, así como un número $\varepsilon^* = \varepsilon^*(\theta^*) > 0$ con la siguiente propiedad:

Para cada $\omega \in \Omega^*$, existe un entero $N(\omega;\theta^*)$ tal que *para todo* $\theta \in B(\theta^*, \varepsilon^*)$

$$f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega);\theta) < f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega);\theta_0), \quad n > N(\omega;\theta^*)$$

Esencialmente, este resultado establece que si $\theta^*$ no coincide con el verladero valor del parámetro $\theta_0$, entonces el maximizador $\theta_n^*$ de la función de verosimilitud basado en $n$ observaciones no se ubica en la bola de centro $\theta^*$ y radio positivo $\varepsilon^* > 0$, a condición de que el número de vectores observados sea suficientemente grande, i.e., $n > N(\omega, \theta^*)$.

**Demostración.** Sea $B(\mathbf{y})$ la función en la Hipótesis 3.2.2, de manera que $E_{\theta_0}[B(\mathbf{Y})] < \infty$ y recuerde que

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

vea al Teorema 2.4.1(i) del Capítulo 2. En este caso, existe un número positivo $\varepsilon^*$ tal que

$$\nu(\theta^*) + \varepsilon^* E_{\theta_0}[B(Y)] < 0 \tag{3.2}$$

de tal manera que

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

A partir de este hecho, la ley de los grandes números establecida en el Teorema 2.2.2 del Capítulo 2 implica que existe un evento $\Omega^*$ con $P[\Omega^*] = 1$ tal que, para todo $\omega \in \Omega^*$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{f(\mathbf{Y}_i(\omega);\theta^*)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) + \varepsilon^* B(\mathbf{Y}_i(\omega))\right] = \nu(\theta^*) + \varepsilon^* E_{\theta_0}[B(Y)].$$

Como el lado derecho es negativo, a partir de la definición de límite se desprende que para cada $\omega \in \Omega^*$ existe un entero positivo $N^* = N^*(\omega)$ tal que

$$\sum_{i=1}^{n}\left[\log\left(\frac{f(\mathbf{Y}_i(\omega);\theta^*)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) + \varepsilon^* B(\mathbf{Y}_i(\omega))\right] < 0, \quad n > N^*(\omega), \quad \omega \in \Omega^*. \tag{3.3}$$

Seleccione ahora un parámetro $\theta$ arbitrario en $B(\theta^*, \varepsilon^*)$, de manera que

$$\|\theta - \theta^*\| < \varepsilon^*. \tag{3.4}$$

Usando la parte (i) de la Hipótesis 3.2.2 se desprende que

$$|\log f(\mathbf{y};\theta) - \log f(\mathbf{y};\theta^*)| \le \|\theta - \theta^*\| B(\mathbf{y}) \le \varepsilon^* B(\mathbf{y}), \quad \mathbf{y} \in \mathcal{Y}$$

Por lo tanto,

$$\begin{aligned}
\log\left(\frac{f(\mathbf{y};\theta)}{f(\mathbf{y};\theta_0)}\right) - \log\left(\frac{f(\mathbf{y};\theta^*)}{f(\mathbf{y};\theta_0)}\right) &= [\log f(\mathbf{y};\theta) - \log f(\mathbf{y};\theta_0)] \\
&\quad - [\log f(\mathbf{y};\theta^*) - \log f(\mathbf{y};\theta_0)] \\
&= \log f(\mathbf{y};\theta) - \log f(\mathbf{y};\theta^*) \\
&\le \varepsilon^* B(\mathbf{y})
\end{aligned}$$

y entonces

$$\log\left(\frac{f(\mathbf{y};\theta)}{f(\mathbf{y};\theta_0)}\right) \leq \log\left(\frac{f(\mathbf{y};\theta^*)}{f(\mathbf{y};\theta_0)}\right) + \varepsilon^* B(\mathbf{y}), \quad \mathbf{y} \in \mathcal{Y}$$

Por lo tanto,

$$\log\left(\frac{f(\mathbf{Y}_i(\omega);\theta)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) \leq \log\left(\frac{f(\mathbf{Y}_i(\omega);\theta^*)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) + \varepsilon^* B(\mathbf{Y}_i(\omega)),$$

de donde se desprende que

$$\sum_{i=1}^{n} \log\left(\frac{f(\mathbf{Y}_i(\omega);\theta)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) \leq \sum_{i=1}^{n} \left[\log\left(\frac{f(\mathbf{Y}_i(\omega);\theta^*)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) + \varepsilon^* B(\mathbf{Y}_i(\omega))\right].$$

Combinando esta desigualdad con (3.3) y recordando que $\theta \in \Theta$ es un parámetro arbitrario que satisface (3.4) se concluye que

$$\sum_{i=1}^{n} \log\left(\frac{f(\mathbf{Y}_i(\omega);\theta)}{f(\mathbf{Y}_i(\omega);\theta_0)}\right) < 0, \quad \omega \in \Omega^*, \quad n > N^*(\omega), \quad \theta \in B(\theta^*, \varepsilon^*). \tag{3.5}$$

Por otro lado, observando que $f_n(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n; \theta) = \prod_{i=1}^{n} f(\mathbf{y}_i; \theta)$ se obtiene

$$\begin{aligned} \log f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta) - \log f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta_0) &= \log\left(\frac{f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta)}{f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta_0)}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{f(\mathbf{y}_i;\theta)}{f(\mathbf{y}_i;\theta_0)}\right) \end{aligned}$$

de tal manera que (3.5) implica que, para cada $\omega \in \Omega^*$ y $\theta \in B(\theta^*, \varepsilon^*)$

$$\log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta) < \log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0), \quad n > N^*(\omega);$$

como $P[\Omega^*] = 1$, esto muestra que la conclusión deseada se satisface con $N(\omega, \theta^*) = N^*(\omega)$ y $\varepsilon^*(\theta^*) = \varepsilon^*$. □

## 3.4 Demostración del Teorema 3.3.1

En esta sección se proporciona una demostración del Teorema 3.3.1. El argumento utiliza la siguiente consecuencia de los Teoremas 2.2.1 y 3.3.2

**Teorema 3.4.1** Sea $\varepsilon > 0$ un número fijo. Bajo las Hipótesis 3.2.1 y 3.2.2 existe un evento $\Omega_\varepsilon$ con las siguiente propiedades:

$$P[\Omega_\varepsilon] = 1$$

y

Para cada $\omega \in \Omega_\varepsilon$ y $\theta \in \Theta \cap B(\theta_0, \varepsilon)^c$ existe un entero $N_\varepsilon(\omega)$ tal que

$$\log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta) < \log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0), \quad n > N_\varepsilon(\omega).$$

Verbalmente, la conclusión de este teorema puede establecerse como sigue: Con probabilidad 1, los maximizadores de $\log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta)$ no se ubican en $\Theta \cap B(\theta_0, \varepsilon)^c$ cuando $n$ es suficientemente grande. Esto indica que dichos maximizadores deben estar a una distancia menor que $\varepsilon$ de $\theta_0$, el cual es el verdadero valor del parámetro.

**Demostración.** Considere el conjunto $\Theta \cap B(\theta_0, \varepsilon)^c$ el cual es un subconjunto compacto de $\Theta$. Para cada $\theta^* \in \Theta \cap B(\theta_0, \varepsilon)^c$, se tiene que $\|\theta^* - \theta_0\| > \varepsilon$; en particular, $\theta^* \neq \theta_0$. Luego, por el Teorema 3.3.2, existe un evento $\Omega(\theta^*)$ con

$$P[\Omega(\theta^*)] = 1 \tag{4.1}$$

así como un número positivo $\varepsilon(\theta^*)$ tal que, para cada $\omega \in \Omega(\theta^*)$ existe un entero $N(\omega, \theta^*)$ con la siguiente propiedad:

$$\begin{aligned} &\text{Si } \|\theta - \theta^*\| < \varepsilon(\theta^*), \text{ entonces para } n > N(\omega, \theta^*) \\ &\log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta) < \log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0). \end{aligned} \tag{4.2}$$

Note ahora que

$$\Theta \cap B(\theta_0, \varepsilon)^c \subset \bigcup_{\theta^* \in \Theta \cap B(\theta_0, \varepsilon)} B(\theta^*, \varepsilon(\theta^*)).$$

Puesto que $\Theta \cap B(\theta_0, \varepsilon)^c$ es un conjunto compacto, por el Teorema 2.2.1 existe un conjnuto finito $\{\theta_1^*, \theta_2^*, \ldots, \theta_r^*\} \subset \Theta \cap B(\theta_0, \varepsilon)$ tal que

$$\Theta \cap B(\theta_0, \varepsilon)^c \subset \bigcup_{i=1}^{r} B(\theta_i^*, \varepsilon(\theta_i^*)). \tag{4.3}$$

Defina el evento $\Omega_\varepsilon$ mediante

$$\Omega_\varepsilon := \bigcap_{i=1}^{r} \Omega(\theta_i^*) \tag{4.4}$$

mientras que

$$N_\varepsilon(\omega) := \max_{i=1,2,\ldots,r} N(\omega, \theta_i^*), \quad \omega \in \Omega_\varepsilon. \tag{4.5}$$

Note ahora que (4.1) y (4.4) implican que $P[\Omega_\varepsilon] = 1$, y que $N_\varepsilon(\omega)$ es finto, pues cada $N(\omega, \theta_i^*)$ lo es. Se mostrará que la conclusión del teorema ocurre con el evento $\Omega_\varepsilon$ y el entero $N_\varepsilon(\omega)$ especificados de esta forma. Sean

$$\theta \in \Theta \cap B(\theta_0)^c \quad \text{y} \quad \omega \in \Omega_\varepsilon$$

arbitrarios y seleccione

$$n > N_\varepsilon(\omega). \tag{4.6}$$

Por la inclusión (4.3), se tiene que $\theta \in B(\theta_i^*, \varepsilon(\theta_i^*))$ para algún $i$ entre 1 y $r$, esto es,

$$\|\theta - \theta_i^*\| < \varepsilon(\theta_i^*)$$

Además, $\omega \in \Omega(\theta_i^*)$, por (4.4), mientras que $n > N(\omega, \theta_i^*)$ por (4.5)y (4.6). Por lo tanto, (4.2) implica que

$$\log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta) < \log f_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega); \theta_0).$$

En resumen: Si $\theta \in \Theta \cap B(\theta_0)^c$, $\omega \in \Omega_\varepsilon$ y $n > N_\varepsilon(\omega)$, entonces la anterior desigualdad ocurre, completando la demostración pues, como ya se ha mencionado, $P[\Omega_\varepsilon] = 1$. □

**Demostración del Teorema 3.3.1.** Para cada entero $m = 1, 2, \ldots$, sea $\Omega_{1/m}$ el evento en el Teorema 3.4.1 correspondiente a $\varepsilon = 1/m$, y defina

$$\Omega^* = \bigcap_{m=1}^{\infty} \Omega_{1/m}. \tag{4.7}$$

En este caso, $\Omega^{*c} = \bigcup_{m=1}^{\infty} \Omega_{1/m}^c$ de manera que $P[\Omega^{*c}] \leq \sum_{m=1}^{\infty} P\left[\Omega_{1/m}^c\right] = 0$, y entonces

$$P[\Omega^*] = 1. \tag{4.8}$$

Se demostrará a continuación que , para cada $\omega \in \Omega^*$, si

$$\hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega))$$

es un maximizador de la función de verosimilitud, entonces

$$\lim_{n\to\infty} \hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) = \theta_0.$$

Para alcanzar este objetivo, sean $\omega \in \Omega^*$ y $\delta > 0$ arbitrarios. Seleccione un entero $m > 0$ tal que $1/m < \delta$ y note que $\omega \in \Omega_{1/m}$, por (4.7). Sea $N_{1/m}(\omega)$ el entero en el Teorema 3.4.1, de tal manera que

$$\begin{aligned} &\text{si } n > N_{1/m}(\omega), \text{ entonces para } \theta \in \Theta \cap B(\theta_0, 1/m)^c \\ &\log f(\mathbf{Y}_1(\omega), \ldots, Y_n(\omega); \theta) < \log f(\mathbf{Y}_1(\omega), \ldots, Y_n(\omega); \theta_0). \end{aligned}$$

Esto significa que para $n > N_{1/m}(\omega)$ la función

$$\theta \mapsto \log f(\mathbf{Y}_1(\omega), \ldots, Y_n(\omega); \theta)$$

no se maximiza en ningún punto de $\theta \in \Theta \cap B(\theta_0, 1/m)^c$, de manera que $\hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) \in \Theta \cap B(\theta_0, 1/m) \subset B(\theta_0, 1/m)$, y entonces

$$\|\hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) - \theta_0\| < \frac{1}{m} < \delta, \quad n > N_{1/m}(\omega).$$

De acuerdo a la definición de límite, esta desigualdad muestra que

$$\lim_{n\to\infty} \hat{\theta}_n(\mathbf{Y}_1(\omega), \ldots, \mathbf{Y}_n(\omega)) = \theta_0;$$

como $\omega \in \Omega^*$ es arbitrario y $P[\Omega^*] = 1$, se cocnluye que $\hat{\theta}_n$ converge a $\theta_0$ con probabilidad 1, completando la demostración. □

# Capítulo 4

# El Argumento Clásico

## 4.1 Introducción

En este capítulo se presenta la argumentación clásica para para establecer la consistencia de los estimadores de verosimilitud máxima en el caso en que el espacio de parámetros $\Theta$ es un subconjunto de $\mathbb{R}$. Como ya se ha mencionado, las condiciones usuales para asegurar la consistencia involucran diferenciabilidad y se establecen a continuación ( Serfling 1988, Greene 2001, Dudewicz, 1989)

**Hipótesis 4.1.1** *La familia* $\{f(\mathbf{y};\theta): \theta \in \Theta\}$ *satisface las siguientes condiciones:*

**R1:** *Para cada* $\mathbf{y} \in \mathcal{Y}$, *la función de densidad* $f(\mathbf{y};\theta)$ *tiene derivadas parciales respecto a* $\theta$ *hasta el tercer orden.*

**R2:** *Dado un parámetro* $\theta_0 \in \Theta$, *existen* $\delta > 0$ *y funciones* $H_1$, $H_2$ *y* $H_3$ *definidas en* $\mathcal{Y}$ *tales que*

$$\left|\frac{\partial^i f(\mathbf{y};\theta)}{\partial \theta^i}\right| \leq H_i(\mathbf{y}), \quad \mathbf{y} \in \mathcal{Y}, \quad \theta \in (\theta_0 - \delta, \theta_0 + \delta), \quad i = 1, 2, 3;$$

**R3:** *Más aún, las funciones $H_i$ satisfacen las siguientes condiciones de integrabilidad:*

$$\int_{\mathcal{Y}} H_i(\mathbf{y})\, d\mathbf{y} < \infty, \quad i = 1, 2;$$

$$0 < E_{\theta_0}\left[\left(\frac{\partial \log f(\mathbf{Y};\theta_0)}{\partial\theta}\right)^2\right] < \infty \quad y \quad 0 < E_{\theta_0}\left[H_3(Y)\right] < \infty.$$

Bajo estas condiciones, el resultado clásico de consistencia sobre estimadores de verosimilitud máxima es el siguiente:

**Teorema 4.1.1** Denote por $\theta_0$ al verdadero valor del parámetro. Bajo la Hipótesis 4.1.1, existe una sucesión de estimadores $\{\hat{\theta}_n = \hat{\theta}_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)\}$ tales que, con probabilidad 1 respecto a $P_{\theta_0}$, las siguientes afirmaciones son válidas:

(i) Para $n$ suficientemente grande, $\hat{\theta}_n$ es una solución para $\theta$ del sistema de ecuaciones de verosimilitud

$$\frac{\partial \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)}{\partial\theta} = 0; \tag{1.1}$$

(ii) $\lim_{n\to\infty} \hat{\theta}_n = \theta_0$, esto es, $\{\hat{\theta}_n\}$ es una sucesión consistente de estimadores del verdadero valor del parametro $\theta$.

En la ecuación (1.1) $\mathcal{L}$ denota el logaritmo de la función de verosimilitud, esto es,

$$\mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) := \log L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \log f_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n; \theta), \tag{1.2}$$

de manera que (1.1) equivale a

$$\frac{\partial L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)}{\partial\theta} = 0$$

y también a

$$\frac{\partial f_n(\mathbf{Y}_1, \ldots, \mathbf{Y}_n; \theta)}{\partial\theta} = 0.$$

Por otro lado, hay al menos dos aspectos en el Teorema 4.1.1 que es conveniente enfatizar. Primero, el teorema segura que, para $n$ suficientemente

grande, el sistema de ecuaciones de verosimilitud tiene una solución $\hat{\theta}_n$, esto es que $\hat{\theta}_n$ es un punto crítico de $\mathcal{L}_n$ como función de $\theta$, pero no garantiza que $\hat{\theta}_n$ sea efectivamente un maximizador de $\mathcal{L}_n$. En los caso prácticos, sin embargo, con frecuencia es posible demostrar que $\mathcal{L}_n$ si tiene un maximizador $\tilde{\theta}_n$ como función de $\theta$, y que el sistema (1.1) tiene solución única; por lo tanto, dicho maximizador debe ser la solución $\hat{\theta}_n$ de (1.1) referida en el teorema, de manera que $\tilde{\theta}_n = \hat{\theta}_n$ converge a $\theta_0$ conforme el tamaño de la muestra aumenta, por la parte (ii) del Teorema 4.1.1. Por otro lado, aún concediendo que la solución del sistema (1.1) corresponda a un maximizador de la función de verosimilitud, su existencia está garantizada sólo en el caso de que las condiciones de diferenciabilidad e integrabilidad en la Hipótesis 4.1.1 se satisfagan. De esta manera, el Teorema 4.1.1 no es aplicable, por ejemplo, en caso de que la muestra provenga de una densidad de Laplace, como en el Ejemplo 3.2.2. En contraste, el Teorema  si es aplicable en este último caso, pues la Hipótesis 3.2.2 se verifica. Una pregunta natural es por qué es necesario suponer las condiciones en la Hipótesis 4.1.1 para poder obtener las conclusiones del Teorema 4.1.1. La respuesta es que el argumnto clásico para demostrar la consistencia de los estimadores de verosimilitud máxima utiliza ciertas ideas analíticas, como desarrollos de Taylor, y las condicones (R1)—(R3) son neceasrias para que dichas expansiones sean válidas. El argumento que se presenta a continuación es una versión más simple del presentado en (Serfling 1988).

## 4.2 Resultados Auxiliares

En esta sección se presentan los resultados preliminares que se utilizan en el enfoque clásico para demostrar el Teorema . Dichos resultados dependen del siguiente resultado básico, conocido como *teorema de convergencia dominada*, el cual establece condiciones bajo las cuales es posible intercambiar las operaciones de límite e integración.

**Teorema 4.2.1** Sea $H(\mathbf{y})$ una función no negativa definida en $\mathcal{Y}$ y suponga

que

$$\int_{\mathcal{Y}} H(\mathbf{y})\, d\mathbf{y} < \infty$$

Si $\{g_n : \mathcal{Y} \to \mathbb{R}\}$ es una sucesión de funciones que satisface

$$|g_n(\mathbf{y})| \leq H(\mathbf{y}), \quad \mathbf{y} \in \mathcal{Y}, \quad n = 1, 2, 3, \ldots \tag{2.1}$$

y que

$$\lim_{n\to\infty} g_n(\mathbf{y}) = g(\mathbf{y}).$$

En este caso las funciones $g_n$ y $g$ son integrables, esto es, $\int_{\mathcal{Y}} |g_n(\mathbf{y})|\, d\mathbf{y} < \infty$ y $\int_{\mathcal{Y}} |g(\mathbf{y})|\, d\mathbf{y} < \infty$, y además

$$\lim_{n\to\infty} \int_{\mathcal{Y}} g_n(\mathbf{y})\, dy = \int_{\mathcal{Y}} \lim_{n\to\infty} g_n(\mathbf{y})\, dy = \int_{\mathcal{Y}} g(\mathbf{y})\, dy. \tag{2.2}$$

Una demostración de este teorema puede encontrarse en Ash (1987), Rudin (1968) o Dudley (2002). Verbalmente, la conclusión de este resultado establece que si la sucesión $\{g_n\}$ esta dominada por una función integrable $H(\mathbf{y})$ (vea (2.1)), entonces es posible intercambiar límite e integración como en (2.2). El siguiente teorema utiliza el anterior resultado para establecer dos importantes consecuencias de la Hipótesis 4.1.1.

**Teorema 4.2.2** Bajo la Hipótesis 4.1.1, sea $\theta \in (\theta_0 - \delta, \theta_0 + \delta)$ arbitrario. En este caso

(i) Las funciones $\partial_\theta f(\mathbf{y};\theta)$ y $\partial_\theta^2 f(\mathbf{y};\theta)$ son integrables sobre $\mathcal{Y}$ y

$$\int_{\mathcal{Y}} \partial_\theta f(\mathbf{y};\theta)\, d\mathbf{y} = 0 = \int_{\mathcal{Y}} \partial_\theta^2 f(\mathbf{y};\theta)\, d\mathbf{y}.$$

(ii) Las esperanzas de $\partial_\theta^2 \log f(\mathbf{Y},\theta)$ y $\partial_\theta \log f(\mathbf{Y},\theta)$ respecto a $P_\theta$ existen y satisfacen

$$E_\theta\left[\partial_\theta^2 \log f(\mathbf{Y},\theta)\right] = -E_\theta\left[\left(\partial_\theta \log f(\mathbf{Y},\theta)\right)^2\right]$$

y

$$E_\theta\left[\partial_\theta \log f(\mathbf{Y},\theta)\right] = 0.$$

**Demostración.** (i) Considere una sucesión $\{\delta_n\}$ de números positivos tal que

$$\lim_{n\to 0} \delta_n = 0,$$

y defina

$$g_n(\mathbf{y}) = \frac{f(\mathbf{y}, \theta + \delta_n) - f(\mathbf{y}, \theta)}{\delta_n}, \quad y \in \mathcal{Y}.$$

Observe que

$$\int_{\mathcal{Y}} g_n(\mathbf{y})\, d\mathbf{y} = \frac{1}{\delta_n}\left[\int_{\mathcal{Y}} f(\mathbf{y}, \theta + \delta_n)\, d\mathbf{y} - \int_{\mathcal{Y}} f(\mathbf{y}, \theta)\, d\mathbf{y}\right] = 0$$

pues tanto $f(\mathbf{y}, \theta + \delta_n)$ y $f(\mathbf{y}, \theta)$ son densidades con soporte $\mathcal{Y}$, mientras que la definición de derivada implica que

$$\lim_{n\to\infty} g_n(\mathbf{y}) = \partial_\theta f(\mathbf{y}; \theta). \tag{2.3}$$

Por otro lado, para $n$ suficientemente grande, $\theta + \delta_n \in (\theta_0 - \delta, \theta_0 + \delta)$, y el teorema del valor medio implica que para cada $\mathbf{y} \in \mathcal{Y}$ existe $\tilde{\delta}_n \in (0, \delta)$ tal que $f(\mathbf{y}, \theta+\delta_n) - f(\mathbf{y}, \theta) = \delta_n \partial_\theta f(\mathbf{y}; \theta+\tilde{\delta}_n)$, de manera que $g_n(\mathbf{y}) = \partial_\theta f(\mathbf{y}; \theta+\tilde{\delta}_n)$, y entonces

$$|g_n(\mathbf{y})| \leq H_1(\mathbf{y}), \quad \mathbf{y} \in \mathcal{Y},$$

por la condición (R2) en la Hipótesis 4.1.1. Puesto que $\int_{\mathcal{Y}} H_1(\mathbf{y}) < \infty$, las tres últimas desigualdades desplegadas permiten concluir, via el Teorema de convergencia dominada, que $\partial_\theta f(\mathbf{y}; \theta)$ es una función integrable, y

$$0 = \lim_{n\to\infty} \int_{\mathcal{Y}} g_n(\mathbf{y})\, d\mathbf{y} = \int_{\mathcal{Y}} \lim_{n\to\infty} g_n(\mathbf{y})\, d\mathbf{y} = \int_{\mathcal{Y}} \partial_\theta f(\mathbf{y}; \theta)\, d\mathbf{y}.$$

La igualdad $\int_{\mathcal{Y}} \partial_\theta^2 f(\mathbf{y}; \theta)\, d\mathbf{y} = 0$ puede establecerse de manera similar.

(ii) Observe que

$$\partial_\theta \log f(\mathbf{y}; \theta) = \frac{\partial_\theta f(\mathbf{y}; \theta)}{f(\mathbf{y}; \theta)} = \partial_\theta f(\mathbf{y}; \theta) \frac{1}{f(\mathbf{y}; \theta)}$$

y entonces

$$\begin{aligned}
\partial_\theta^2 \log f(\mathbf{y};\theta) &= \partial_\theta \left[\partial_\theta f(\mathbf{y};\theta) \frac{1}{f(\mathbf{y};\theta)}\right] \\
&= \partial_\theta \left[\partial_\theta f(\mathbf{y};\theta)\right] \frac{1}{f(\mathbf{y};\theta)} + \partial_\theta f(\mathbf{y};\theta) \partial_\theta \left[\frac{1}{f(\mathbf{y};\theta)}\right] \\
&= \partial_\theta^2 f(\mathbf{y};\theta) \frac{1}{f(\mathbf{y};\theta)} + \partial_\theta f(\mathbf{y};\theta) \frac{\partial_\theta f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)^2} \\
&= \frac{\partial_\theta^2 f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} - \frac{\partial_\theta f(\mathbf{y};\theta)^2}{f(\mathbf{y};\theta)^2} \\
&= \frac{\partial_\theta^2 f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} - \left(\frac{\partial_\theta f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\right)^2 \\
&= \frac{\partial_\theta^2 f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} - \left(\partial_\theta \log f(\mathbf{y};\theta)\right)^2
\end{aligned}$$

Por lo tanto,

$$\partial_\theta^2 \log f(\mathbf{y};\theta)\, f(\mathbf{y};\theta) = \partial_\theta^2 f(\mathbf{y};\theta) - \left(\partial_\theta \log f(\mathbf{y};\theta)\right)^2 f(\mathbf{y};\theta)$$

y entonces

$$\begin{aligned}
&\int_{\mathcal{Y}} \partial_\theta^2 \log f(\mathbf{y};\theta)\, f(\mathbf{y};\theta)\, d\mathbf{y} \\
&\qquad = \int_{\mathcal{Y}} \partial_\theta^2 f(\mathbf{y};\theta)\, d\mathbf{y} - \int_{\mathcal{Y}} \left(\partial_\theta \log f(\mathbf{y};\theta)\right)^2 f(\mathbf{y};\theta)\, d\mathbf{y}.
\end{aligned}$$

y usando la igualdad $\int_{\mathcal{Y}} \partial_\theta^2 f(\mathbf{y};\theta)\, d\mathbf{y} = 0$ establecida en la primera parte del teorema se desprende que

$$\int_{\mathcal{Y}} \partial_\theta^2 \log f(\mathbf{y};\theta)\, f(\mathbf{y};\theta)\, d\mathbf{y} = -\int_{\mathcal{Y}} \left(\partial_\theta \log f(\mathbf{y};\theta)\right)^2 f(\mathbf{y};\theta)\, d\mathbf{y},$$

lo cual equivale a

$$E_\theta\left[\partial_\theta^2 \log f(Y;\theta)\right] = -E_\theta\left[\left(\partial_\theta \log f(Y;\theta)\right)^2\right].$$

Para concluir, observe que la igualdad $\int_{\mathcal{Y}} \partial_\theta f(\mathbf{y};\theta)\, d\mathbf{y} = 0$ establecida en la parte (i), equivale a

$$\int_{\mathcal{Y}} \partial_\theta \log f(\mathbf{y};\theta)\, d\mathbf{y} = \int_{\mathcal{Y}} \frac{\partial_\theta f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} f(\mathbf{y};\theta)\, d\mathbf{y} = 0,$$

esto es, $E_\theta\left[\partial_\theta \log f(Y;\theta)\right] = 0$, concluyendo la demostración. □

## 4.3 Cotas Para la ecuación de Verosimilitud

En esta sección se obtienen cotas para la derivada parcial respecto a $\theta$ del logaritmo de la función de verosimilitud alrededor del verdadero valor del parámetro $\theta_0$. Dichas cotas son funciones cuadráticas, cuyas raíces pueden encotrarse fácilmente, y serán el instrumento básico para encontrar soluciones de la ecuación de verosimilitud (1.1). Como en Serfling (1988), la idea fundamental es obtener el desarrollo de Taylor de la función $\partial_\theta$ alrededor de $\theta_0$, pero despues de ese punto el argumento que sigue es completamente distinto y más simple que el usualmente aplicado. Como punto de partida, note que después de observar los datos $Y_1, Y_2, \ldots, Y_n$, la verosimilitud del parámetro $\theta$ es

$$L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \prod_{i=1}^{n} f(\mathbf{Y}_i; \theta)$$

de manera que

$$\mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \log L_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \sum_{i=1}^{n} \log f(\mathbf{Y}_i; \theta)$$

y esta funcón tiene, por lo menos, hasta tercera derivada respecto a $\theta$, por la Hipótesis 4.1.1. Por lo tanto

$$\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \sum_{i=1}^{n} \partial_\theta \log f(\mathbf{Y}_i; \theta) \tag{3.1}$$

tiene por lomenos dos derivadas respecto a $\theta$. A partir de este hecho se desprende que $\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ tiene un desarrollo de Taylor alrededor del verdadero valor del parámetro $\theta_0$:

$$\begin{aligned} &\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \\ &\quad = \partial_\theta \mathcal{L}_n(\theta_0; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) + \partial_\theta^2 \mathcal{L}_n(\theta_0; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)(\theta - \theta_0) \\ &\qquad + \partial_\theta^2 \mathcal{L}_n(\tilde{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)(\theta - \theta_0)^2 \end{aligned} \tag{3.2}$$

donde $\tilde{\theta}$ es un punto entre $\theta$ y $\theta_0$. Las cotas para $\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ en el siguiente lema serán útiles en el desarrollo subsecuente.

**Lema 4.3.1** Sean $\theta_0$ el verdadero valor del parámetro y $\delta > 0$ como en la Hipótesis 4.1.1. Para $\theta \in (\theta_0 - \delta, \theta_0 + \delta)$,

$$\frac{1}{n}\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \leq A_n - B_n(\theta - \theta_0) + C_n(\theta - \theta_0)^2 \tag{3.3}$$

y

$$\frac{1}{n}\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \geq A_n - B_n(\theta - \theta_0) - C_n(\theta - \theta_0)^2 \tag{3.4}$$

donde

$$A_n = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \log f(\theta_0; \mathbf{Y}_i); \tag{3.5}$$

$$B_n = -\frac{1}{n}\sum_{i=1}^{n} \partial_\theta^2 \log f(\theta_0; \mathbf{Y}_i), \tag{3.6}$$

y

$$C_n = \frac{1}{n}\sum_{i=1}^{n} H_3(\mathbf{Y}_i), \tag{3.7}$$

donde $H_3(\mathbf{y})$ es la cota para $\partial_\theta^3 \log f(\mathbf{y}; \theta)$ postulada en la parte (R2) de la Hipótesis 4.1.1.

**Demostración.** A partir de (3.1) se desprende que

$$\partial_\theta \mathcal{L}_n(\theta_0; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \sum_{i=1}^{n} \partial_\theta \log f(\theta_0; \mathbf{Y}_i) = nA_n;$$

$$\partial_\theta^2 \mathcal{L}_n(\theta_0; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \sum_{i=1}^{n} \partial_\theta^2 \log f(\theta_0; \mathbf{Y}_i) = -nB_n,$$

donde las igualdades en la extrema derecha se desprenden de la especificación de $A_n$ y $B_n$ en (3.5) y (3.6), respectivamente, y entonces (3.2) implica que

$$\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = nA_n - nB_n(\theta - \theta_0) + \partial_\theta^2 \mathcal{L}_n(\tilde{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)(\theta - \theta_0)^2. \tag{3.8}$$

Por otro lado, $\partial_\theta^3 \mathcal{L}_n(\tilde{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \sum_{i=1}^{n} \partial_\theta^3 \log f(\tilde{\theta}; \mathbf{Y}_i)$; como $\theta$ pertenence a $(\theta_0 - \delta, \theta_0 + \delta)$ y $\tilde{\theta}$ se encuentra entre $\theta$ y $\theta_0$, se tiene que $\tilde{\theta} \in$

$(\theta_0 - \delta, \theta_0 + \delta)$, de manera que, a partir del supuesto (R2) en la Hipótesis 4.1.1, se obtiene que

$$\left|\partial_\theta^3 \log f(\tilde{\theta}; \mathbf{Y}_i)\right| \leq H_3(\mathbf{Y}_i)$$

y entonces

$$\left|\partial_\theta^3 \mathcal{L}_n(\tilde{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)\right| \leq \sum_{i=1}^{n} \left|\partial_\theta^3 \log f(\tilde{\theta}; \mathbf{Y}_i)\right| \leq \sum_{i=1}^{n} H_3(\mathbf{Y}_i) = nC_n;$$

donde (3.7) se usó para estableer la igualdad. Combinando este hecho con (3.8) se desprende que

$$\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \leq nA_n - nB_n(\theta - \theta_0) + nC_n(\theta - \theta_0)^2$$

y

$$\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \geq nA_n - nB_n(\theta - \theta_0) - nC_n(\theta - \theta_0)^2,$$

relaciones que son equivalentes a (3.3) y (3.4), respectivamente. □

## 4.4 Raíces de las Cotas Cuadráticas

En esta sección se estudian las cotas cuadráticas para $\partial_\theta \mathcal{L}_n(\theta; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ establecidas en el Lema 4.3.1. Por conveniencia, escriba

$$\begin{aligned} U_{1\,n}(\theta) &= A_n - B_n(\theta - \theta_0) + C_n(\theta - \theta_0)^2; \\ U_{0\,n}(\theta) &= A_n - B_n(\theta - \theta_0) - C_n(\theta - \theta_0)^2. \end{aligned} \tag{4.1}$$

El siguiente objetivo es mostrar que, con probabilidad 1, cuando $n$ es suficientemente grande estas funciones cuadráticas tiene raíces $\theta_{1\,n}$ y $\theta_{0\,n}$ que convergen a $\theta_0$. El primer paso en esta dirección se establece a continuación y depende de la fórmula cuadrática.

**Lema 4.4.1** Suponga que

$$B_n^2 - 4|A_n C_n| > 0, \quad \text{y} \quad C_n > 0. \tag{4.2}$$

y defina $\theta_{0\,n}$ y $\theta_{1\,n}$ mediante

$$\begin{aligned}\theta_{1\,n} &= \theta_0 + \frac{B_n - \sqrt{B_n^2 - 4A_nC_n}}{2C_n} \\ \theta_{0\,n} &= \theta_0 + \frac{B_n - \sqrt{B_n^2 + 4A_nC_n}}{-2C_n}\end{aligned} \tag{4.3}$$

Entonces

$$U_i(\theta_{1\,n}) = 0 = U_0(\theta_{0\,n}) \tag{4.4}$$

**Demostración.** Suponga que (4.2) es válida. Observe ahora que la ecuación $U_{1\,n}(\theta) = 0$ es cuadrática en $z = \theta - \theta_0$, y que su discriminante es

$$B_n^2 - 4A_n(C_n) \geq B_n^2 - 4|A_nC_n| > 0$$

Usando la fórmula cuadrática, la ecuación $U_{1\,n}(\theta) = 0$ es equivalente a

$$\theta - \theta_0 = \frac{B_n \pm \sqrt{B_n^2 - 4A_nC_n}}{2C_n};$$

seleccionando el signo menos se obtiene que

$$\theta = \theta_0 + \frac{B_n - \sqrt{B_n^2 - 4A_nC_n}}{2C_n} = \theta_{1\,n}$$

satisface $U_{1\,n}(\theta_{1\,n}) = 0$. Considere ahora la ecuación $U_{0\,n}(\theta) = 0$, la cual de nueva cuenta es cuadrática en $z = \theta - \theta_0$. De acuerdo a (4.1), su discriminante es

$$B_n^2 + 4A_n(C_n) \geq B_n^2 - 4|A_nC_n| > 0$$

y la fórmula cuadrática permite concluir que $U_{1\,n}(\theta) = 0$ equivale a

$$\theta - \theta_0 = \frac{B_n \pm \sqrt{B_n^2 + 4A_nC_n}}{-2C_n},$$

y al seleccionar el signo menos se obtiene que

$$\theta = \theta_0 + \frac{B_n - \sqrt{B_n^2 - 4A_nC_n}}{-2C_n} = \theta_{0\,n}$$

satisface $U_{0\,n}(\theta_{0\,n}) = 0$, concluyendo la demostración. □

EL siguiente resultado muestra que, con probabilidad 1, las condiciones en (4.2) se satisfacen siempre que $n$ es suficientemente grande.

**Lema 4.4.2** Sea $\theta_0$ el verdadero valor del parámetro y suponga que la Hipótesis 4.1.1 es válida. En este caso, las siguientes convergencias ocurren con probabilidad 1:

(i) $\lim_{n\to\infty} A_n = 0$;

(ii) $\lim_{n\to\infty} B_n = -E_{\theta_0}\left[\partial_\theta^2 \log f(\mathbf{Y};\theta)\right] =: B > 0$

(iii) $\lim_{n\to\infty} C_n = -E_{\theta_0}\left[H_3(\mathbf{Y})\right] =: C > 0$

(iv) $\lim_{n\to\infty} B_n^2 - 4|A_nC_n| = B^2 > 0$

**Demostración.** (i) A partir de (3.5) se desprende que $A_n$ es la media muestral de las variables aleatorias $\partial_\theta \log f(\mathbf{Y}_i;\theta_0)$, $i = 1, 2, \ldots, n$ las cuales son iid con esperanza 0; vea el Teorema 4.2.2(ii). Por lo tanto, la ley de los grandes números establecida en el Teorema 2.2.1 del Capítulo 2 implica que $\lim_{n\to\infty} A_n = 0$ ocurre con probabilidad 1.

(ii) De la especificación de $B_n$ en (3.6) se desprende que $B_n$ es el promedio de las variables aleatorias $-\partial_\theta^2 \log f(\mathbf{Y}_i;\theta_0)$, $i = 1, 2, \ldots, n$; estas variables son independientes con la misma distribución y, más aún, de acuerdo al Teorema 4.2.2(ii) su valor esperado es

$$B := -E_{\theta_0}\left[\partial_\theta^2 \log f(\mathbf{Y}_i;\theta_0)\right] = E_{\theta_0}\left[\left(\partial_\theta \log f(\mathbf{Y}_i;\theta_0)\right)^2\right] > 0$$

donde la desigualdad se desprende de la condición (R3) en la Hipóteisi 4.1.1. En consecuencia, la ley de los grandes números permite concluir que la convergencia $\lim_{n\to\infty} B_n = B$ es válida con probabilidad 1.

(iii) $C_n$ es la media muestral de las variables aleatorias $H_3(\mathbf{Y}_i)$, las cuales tienen esperanza finita y positiva $C := E_{\theta_0}[H_3(\mathbf{Y})]$, por la condición (R3) en la Hipótesis 4.1.1. Por lo tanto, $\lim_{n\to\infty} C_n = C > 0$, por la ley de los grandes números.

(iv) A partir de las partes (i)–(iii) se desprende que, con probabilidad 1, $\lim_{n\to\infty} B_n^2 - 4|A_nC_n| = B^2 - 4|0C| = B^2 > 0$. □

La última etapa antes de la demostración del Teorema 4.1.1 es el siguiente resultado, el cual muestra que las soluciones $\theta_{i\,n}$ de las ecuaciones cuadráticas $U_{i\,n}(\theta) = 0$ convergen a $\theta_0$ con probabilidad 1.

**Lema 4.4.3** Con probabilidad 1 las siguientes afirmaciones son válidas:

(i) Para $n$ suficientemente grande la condición (4.2) se satisface. Por lo tanto las soluciones $\theta_{1\,n}$ y $\theta_{0\,n}$ en el (4.3) están definidas.

(ii) Las siguientes convergencias ocurren:

$$\lim_{n\to\infty} \theta_{1\,n} = \theta_0$$

y

$$\lim_{n\to\infty} \theta_{0\,n} = \theta_0$$

**Demostración.** (i) Por el lema anterior, las convergencias $\lim_{n\to\infty} C_n = C > 0$ y $\lim_{n\to\infty} B_n^2 - 4|A_nC_n| = B^2 > 0$ ocurren con probabilidad 1, de manera que la definición de límite permite concluir que para $n$ suficientemente grande, $B_n^2 - 4|A_nC_n| > 0$ y $C_n > 0$, que es precisamente la condición (4.2), y en este caso $\theta_{1\,n}$ y $\theta_{0\,n}$ están bien definidas.

(ii) A partir de las fórmulas (4.3) se desprende que

$$\begin{aligned}\lim_{n\to\infty} \theta_{1\,n} &= \theta_0 + \lim_{n\to\infty} \frac{B_n - \sqrt{B_n^2 - 4A_nC_n}}{2C_n} \\ &= \theta_0 + \lim_{n\to\infty} \frac{B - \sqrt{B^2}}{2C}\end{aligned}$$

donde la segunda igualdad se debe al Lema 4.4.2. Como $B > 0$, se tiene que $\sqrt{B^2} = B$,, y entonce la anterior igualdad desplegada se reduce a $\lim_{n\to\infty} \theta_{1\,n} = \theta_0$. La convergancia $\lim_{n\to\infty} \theta_{0\,n} = \theta_0$ se establece de manera similar. □

## 4.5 Demostración del Teorema 4.1.1

Despues de los preliminares técnicos en las secciones precedentes, a continuación se proporcionará una demostración del resultado clásico de existencia de soluciones de la ecuación de verosimilitud (1.1) y de su convergencia hacia el verdadero valor del parámetro $\theta_0$. Sean $\theta_{0\,n}$ y $\theta_{1\,n}$ como en (4.3), valores que, con probabilidad 1, están bien definidos para $n$ suficientemente grande. Sea $\delta > 0$ como en la condición (R2) de la Hipótesis 4.1.1. Como $\lim_{n\to\infty} \theta_{i\,n} = \theta_0 \in (\theta_0 - \delta, \theta_0 + \delta)$, $i = 1, 2$, por el Lemma 4.4.3. Por lo tanto, con probabilidad 1,

$$\theta_{0\,n}, \theta_{1\,n} \in (\theta_0 - \delta, \theta_0 + \delta)$$

para $n$ suficientemente grande. Por lo tanto, las desigualdades (3.3) y (3.4), válidas para $\theta \in (\theta_0 - \delta, \theta_0 + \delta)$ permiten concluir que

$$\begin{aligned}\frac{1}{n}\partial_\theta \mathcal{L}_n(\theta_{1\,n}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) &\leq A_n - B_n(\theta_{1\,n} - \theta_0) + C_n(\theta_{1\,n} - \theta_0)^2 \\ &= U_{1\,n}(\theta_{1\,n}) = 0\end{aligned}$$

donde la prinera igualdad se debe a la definición de $U_{1\,n}$ (vea (4.1)) y la segunda a la especificación de $\theta_{1\,n}$. Similarmente puede demostrarse que

$$\begin{aligned}\frac{1}{n}\partial_\theta \mathcal{L}_n(\theta_{1\,n}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) &\geq A_n - B_n(\theta_{1\,n} - \theta_0) - C_n(\theta_{1\,n} - \theta_0)^2 \\ &= U_{0\,n}(\theta_{0\,n}) = 0.\end{aligned}$$

Considere ahora el intervalo $I_n$ que une a los puntos $\theta_{0\,n}$ y $\theta_{1\,n}$, el cual está contenido en $(\theta_0 - \delta, \theta_0 + \delta)$, y observe qu la funcón $x \mapsto \partial_\theta \mathcal{L}_n(x; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ es continua en $I_n$, pues es derivable, y que de acuerdo a las desigualdades anteriores, esta función es no negativa en un extremo del intervalo y no positiva en el otro. Por lo tanto, la función $x \mapsto \partial_\theta \mathcal{L}_n(x; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ se anula en algún punto de $I_n$, por la propiedad del valor intermedio,, esto es, existe $\hat{\theta}_n \in I_n$ tal que

$$\partial_\theta \mathcal{L}_n(\hat{\theta}_n; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = 0.$$

demostrando la existencia de soluciones a la ecuación de verosimilitud .Para conluir observe que la inclusión $\hat{\theta}_n \in I_n$, y el hecho de que los extremos

del intervalo $I_n$, a saber, $\theta_{0\,n}$ y $\theta_{1\,n}$ convergen con probabilidad 1 hacia $\theta_0$, implica que

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0$$

casi seguramente, completando la demostración del Teorema 4.1.1 $\square$

# Literatura Citada

[1]. R. Ash (1987) *Real Analysis and Probability*, *Academic Press*, New York.

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

[3]. E. Dudewicz y N. Mishra (1989) *Modern Mathematical Statistics*, *Jhon WIley & Sons*, New York.

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

[5]. J. Dugundji (970) Topology, *Allyn & Bacon*, Boston.

[6]. W. Fulks, *Cálculo Avanzado*, *Limusa*, México D.F.

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

[8]. W. E. Griffiths, R. Carter-Hill y George G. Judge (1997) *Learning and Practicing Econometrics*, *Jhon Willey & Sons*, New York.

[9]. E. L. Lehmann (2001) *Testing Statistical Hypotheses*, *Jhon Willey & Sons*, New York.

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

[11]. J. W. Hardin (2002) The robust variance estimator for two-stage models, *The Stata Journal*, **3**, pp. 253266.

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

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

[14]. C. R. Rao (2002) *Linear Statistical Inference and Its Application*, *Jhon Willey*, New York.

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

[16]. R. J. Serfling (1988) *Approximation Theorems of Mathematical Statistics*, *Jhon Willey*, New York.

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