UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO

INSTITUTO DE FÍSICA Y MATEMÁTICAS

"Comportamiento asintótico de la prueba de razón de verosimilitudes"

TESINA

QUE PARA OBTENER EL TÍTULO DE:

MAESTRO EN CIENCIAS MATEMÁTICAS

PRESENTA:

ILIA PAVLOVICH NAUMKIN KAIKINA

ASESOR:

DR. EUGENIO BALANZARIO GUTIÉRREZ

Centro de Ciencias Matemáticas, Universidad Nacional Autónoma de México

---

MORELIA, MICHOACÁN - SEPTIEMBRE de 2018

ÍNDICE GENERAL

Resumen

En estadística, una prueba de razón de verosimilitudes es una prueba de hipótesis utilizada para comparar la bondad de ajuste de dos modelos estadísticos: un modelo nulo frente a un modelo alternativo. La prueba se basa en la razón de verosimilitudes, que expresa cuántas veces más probable es que los datos estén bajo un modelo que bajo el otro. La estadística de prueba depende de una estadística mínima suficiente solamente. Esto es inmediato debido a su definición como un cociente y la caracterización de la suficiencia por el teorema de la factorización. Esta razón de verosimilitudes, o su logaritmo de manera equivalente, se puede usar para decidir si se rechaza o no el modelo nulo. Cuando se usa el logaritmo de la razón de verosimilitudes, la estadística se conoce como una estadística de razón de verosimilitudes logarítmica, y la distribución de probabilidad de esta estadística de prueba, suponiendo que el modelo nulo es verdadero, se puede aproximar usando el teorema de Wilks. En este trabajo vamos a dar una demostración completa y rigurosa del famoso teorema de Wilks, el cual afirma que si la hipótesis nula es verdadera, entonces la distribución de la razon de verosimilitudes λ_n converge cuando n tiende a infinito a una distribución ji-cuadrada con r grados de libertad, no dependiendo del valor verdadero $\theta = \theta_0 \in \Theta_0$. Por lo tanto, la hipótesis nula sería rechazado si λ_n es demasiado grande en términos de la distribución tabulada ji-cuadrada. Esto significa que para una gran variedad de hipótesis, un profesional puede calcular la razón de verosimilitudes λ_n para los datos y comparar λ_n con valor ji-cuadrada correspondiente a la significación estadística deseada como una prueba estadística aproximada.

Palabras clave:
- Razón de verosimilitudes
- Teorema de Wilks
- Distribución ji-cuadrada
- Hipótesis nula
- Distribución asintótica

Abstract

In statistics, a likelihood ratio test is a hypothesis test used to compare two statistical models: a null model versus an alternative model. The test is based on the likelihood ratio, which expresses how many times more likely the data is under one model than under the other. This likelihood ratio, or its equiva-lent logarithm, can be used to decide whether or not to reject the null model. When the logarithm of the likelihood ratio is used, the statistic is known as a logarithmic likelihood ratio statistic, and the probability distribution of this test statistic, assuming that the null model is true, can be approximated using the Wilks theorem. In this work we will give a complete and rigorous demonstration of the famous Wilks theorem, which states that if the null hypothesis is true, then the distribution of the likelihood ratio $_n\lambda$converges when $n$tends to infinity to a chi-square distribution with $r$degrees of freedom, not depending on the true value θ. Therefore, the null hypothesis would be rejected if $_n\lambda$is too large in terms of the chi-square distribution.

1. Introducción

En estadística, una prueba de razón de verosimilitudes es una prueba de hipótesis utilizada para comparar la bondad de ajuste de dos modelos estadísticos: un modelo nulo frente a un modelo alternativo. La prueba se basa en la razón de verosimilitudes, que expresa cuántas veces más probable es que los datos estén bajo un modelo que bajo el otro. Esta razón de verosimilitudes, o su logaritmo de manera equivalente, se puede usar para decidir si se rechaza o no el modelo nulo (vease Subseccion 2.2). Cuando se usa el logaritmo de la razón de verosimilitudes, la estadística se conoce como una estadística de razón de verosimilitudes logarítmica, y la distribución de probabilidad de esta estadística de prueba, suponiendo que el modelo nulo es verdadero, se puede aproximar usando el teorema de Wilks. Samuel S. Wilks publicó por primera vez su teorema en un artículo [12], y luego lo expuso en su libro [13] (sección 13.8). Chernoff [2] dio otra prueba. Van der Vaart [11] (capítulo 16) ofrece una exposición más reciente, donde sugiere que la prueba original de Wilks no era rigurosa. En este trabajo vamos a dar una demostración completa y rigurosa del famoso teorema de Wilks.

Sea un vector aleatorio $(X_1, X_2, ..., X_n) \in \mathbb{R}^n$ de variables independientes e idÃ©nticamente distribuidas (i.i.d.) con función de distribución conjunta $f(x\,|\theta\,)$ en donde θ es un parámetro vectorial de dimensión k que toma valores en Θ, una región abierta de $\mathbb{R}^k$: $\theta \in \Theta \subset \mathbb{R}^k$. Si θ_0 es el valor verdadero de θ en la población Θ_0, un subconjunto de Θ, se plantean las siguientes hipótesis estadísticas: la hipótesis nula $H_0 : \theta \in \Theta_0$ contra la hipótesis alternativa $H_1 : \theta \in \Theta \setminus \Theta_0$. Se trata de decidir si se acepta o si se rechaza la hipótesis nula H_0.

Se define la razón de verosimilitudes

$$\lambda_n = \frac{L_n\left(\theta_n^*\right)}{L_n\left(\widehat{\theta}_n\right)}, \tag{1.1}$$

donde

$$L_n\left(\theta_n^*\right) = \sup_{\theta\in\Theta_0} \prod_{j=1}^{n} f\left(x_j\,|\theta\,\right),$$

$$L_n\left(\widehat{\theta}_n\right) = \sup_{\theta\in\Theta} \prod_{j=1}^{n} f\left(x_j\,|\theta\,\right),$$

y donde θ_n^* es el estimador de máxima verosimilitud (M.V.) de θ sobre Θ_0 y análogamente $\widehat{\theta}_n$ es el estimador de máxima verosimilitud (M.V.) de θ sobre Θ (vease sección 2.1 para más detalles).

Supongamos que se cumplen las siguientes condiciones:

(1) Θ es un conjunto abierto en $\mathbb{R}^k$.

(2) Las segundas derivadas con respecto a θ de la función $f(x\,|\theta\,)$ existen y son continuas. Además se puede intercambiar los simbolos de la integral y

las derivadas en la integral $\int f(x\,|\theta)d\nu(x)$ en donde $d\nu(x)$ denota la medida de Lebesque, o una medida de conteo.

(3) Existe la función $K(x)$ tal que la esperanza $\mathbb{E}_{\theta_0}(K(x))$ es acotada y cada componente de $\dot{\Psi}(x,\theta)$ está acotada en valor absoluto por la función $K(x)$ en una vecindad del punto θ_0, donde

$$\Psi(x,\theta) = \left(\frac{\partial}{\partial\theta}\log f(x\,|\theta)\right)^T \text{ es un vector de dimensión } k$$

$$\dot{\Psi}(x,\theta) = \frac{\partial^2}{\partial\theta^2}\log f(x\,|\theta) \text{ es una matriz } k\times k$$

(4) La cantidad de información de Fisher $J(\theta_0) = -\mathbb{E}_{\theta_0}\left(\dot{\Psi}(x,\theta_0)\right)$ está definida positivamente (véase subsección 2.5 para más detalles).

(5) Si $f(x\,|\theta) = f(x\,|\theta_0)$ c.s. respecto a la medida $d\nu(x)$, entonces $\theta = \theta_0$.

Ahora enunciamos el teorema de Wilks.

Teorema 1. *(Teorema de Wilks) Supongamos que se cumplen las condiciónes (1)-(5). Sea la hipótesis*

$$H_0 : \theta^1 = \theta^2 = ... = \theta^r = 0, 1 \le r \le k.$$

Supongamos que el valor verdadero θ_0 satisface la hipótesis H_0. Entonces

$$-2\log\lambda_n \xrightarrow{d} \chi^2_r,$$

cuando $n\to\infty$.

Comentarios. 1) La condición (1) del teorema (que significa la "linealidad local"de las hipótesis) es esencial para la aproximación ji-cuadrada, que a veces falla en una serie de ejemplos simples. Un conjunto abierto es ciertamente localmente lineal en cada uno de sus puntos, y lo también es un subconjunto relativamente abierto de un subespacio afín. Por otro lado, cuando se prueba una hipótesis unilateral $H_0 : \mu(\theta) \le 0$, o una hipótesis de la forma $H_0 : |\theta| \le 1$, entonces no tenemos condiciones de linealidad en los puntos de frontera. En ese caso, la distribución asintótica de la estadística de razón de verosimilitudes no es ji-cuadrada, sino la distribución de un cierto funciónal de un vector gaussiano.

2) Wilks encontró que si la hipótesis H_0 es verdadera, entonces la distribución de λ_n converge cuando $n\to\infty$ a una distribución ji-cuadrada con r grados de libertad, no dependiendo del valor verdadero $\theta = \theta_0 \in \Theta_0$. Por lo tanto, H_0 sería rechazado si λ_n es demasiado grande en términos de la distribución tabulada ji-cuadrada. Esto significa que para una gran variedad de hipótesis, un profesional puede calcular la razón de verosimilitudes λ_n para los datos y comparar λ_n con valor χ^2_r correspondiente a la significación estadística deseada como una prueba estadística aproximada.

El resto del trabajo está organizado de la siguiente manera. En la sección 2 vamos a dar algunas definiciones, que se usan en el trabajo. La sección 3 está dedicada a las demostraciones de algunos resultados auxiliares, que

vamos a aplicar en la demostración del Teorema de Wilks. Finalmente el resultado principal del trabajo está demostrado en la sección 4.

2. Definiciones

En esta sección se verán algunos resultados matemáticos o estadísticos útiles para las secciones siguientes.

2.1. Método de máxima verosimilitud. Introducimos el concepto de estimación de máxima verosimilitud. En esencia, el método de estimación por máxima verosimilitud, selecciona como estimador a aquél valor del parámetro θ que tiene la propiedad de maximizar el valor de la probabilidad de la muestra aleatoria observada. En otras palabras, el método de máxima verosimilitud consiste en encontrar el valor del parámetro θ que maximiza la función de verosimilitud. Ahora bien, sea $x_1, x_2, ..., x_n$ una muestra aleatoria de una distribución con función de densidad (ó probabilidad) $f(x|\theta)$, y definimos $L_n(\theta)$ como la función de verosimilitud

$$L_n(\theta) = \prod_{j=1}^{n} f(x_j|\theta)$$

y

$$l_n(\theta) = \log(L_n(\theta)) = \sum_{j=1}^{n} \log f(x_j|\theta).$$

Supongamos que existen las derivadas parciales $\frac{\partial}{\partial\theta} f(x_j|\theta)$. Entonces el estimador de máxima verosimilitud (EMV) $\widehat{\theta}_n$ es solución de las ecuaciones de verosimilitud

$$\dot{l}_n(\theta) = \frac{\partial}{\partial\theta} \log(L_n(\theta)) = \sum_{j=1}^{n} \frac{\partial}{\partial\theta} \log f(x_j|\theta) = 0.$$

El método de máxima verosimilitud (MV) tiene la propiedad (deseable) de proporcionar estimadores que son funciónes de estadísticos suficientes, siempre y cuando el estimador MV sea único. Además, el método MV proporciona el estimador eficiente, si es que existe. Sin embargo, los estimadores MV son generalmente sesgados. El procedimiento para obtener este tipo de estimadores es (relativamente) directo. Debido a la naturaleza de la función de verosimilitud se escoge, por lo común, maximizar el logaritmo natural de $L(\theta)$. En muchas ocasiones es más fácil obtener el estimador MV maximizando $\log L(\theta)$ que $L(\theta)$.

2.2. Pruebas de Hipótesis. Sea un vector aleatorio $(X_1, X_2, ..., X_n) \in \mathbb{R}^n$ de variables i.i.d. con función de distribución conjunta $f(x|\theta)$ en donde θ es un parámetro vectorial de dimension k que toma valores en Θ, una región abierta de $\mathbb{R}^k$. Recordemos que si θ_0 es el valor verdadero de θ en la población Θ_0, un subconjunto de Θ, se plantean las hipótesis estadísticas: la hipótesis nula $H_0 : \theta_0 \in \Theta_0$ contra la hipótesis alternativa $H_1 : \theta \in \Theta \setminus \Theta_0$.

Se trata de decidir si se acepta o si se rechaza la hipótesis nula H_0. Para resolver esta pregunta, se necesita una regla de decisión. Cualquier regla de decisión debería tratar de minimizar los errores de decisión. Si δ es la regla de decisión adoptada y $\alpha(\delta)$ la probabilidad de equivocarse cuando la hipótesis nula es cierta y $\beta(\delta)$ la probabilidad de equivocarse cuando la hipótesis alternativa es cierta, uno buscará minimizar ambas probabilidades de error. Dada una hipótesis nula H_0, $\alpha(\delta)$ es la probabilidad condiciónal de rechazar la hipótesis H_0 con la regla δ cuando H_0 es cierta. Ahora bien la regla δ se basa en los valores muestrales: si la muestra es de tamaño n y los valores muestrales en $\mathbb{R}^n$, una regla de decisión δ consiste en dividir el dominio $\mathbb{R}^n$ del conjunto de todas las muestras de tamaño n en dos partes disjuntas: la parte W_n en donde se rechaza la hipótesis nula H_0 y la parte $\overline{W_n}$ en donde no se rechaza H_0. La parte W_n se llama región de rechazo de H_0 o región crítica de la prueba. Como la región crítica de la prueba es aquella en donde se rechaza H_0, debería tomarse en cuenta la hipótesis alternativa. Una regla de decisión consiste entonces en determinar la región crítica de la prueba en función de las dos hipótesis. Describimos dos tipos de pruebas (véase [1]):

Prueba uniformemente más potente (UMP). Queremos construir una región crítica más potente con hipótesis no simples. Se dice que una prueba es UMP (uniformemente más potente) cuando existe una región crítica óptima común para todo valor de la hipótesis alternativa H_1.

Sea la hipótesis nula $H_0 : \theta = \theta_0$ y la hipótesis alternativa $H_1 : \theta > \theta_0$ (o $H_1 : \theta \neq \theta_0$). La región crítica óptima de nivel de significación α no cambia para todo $\theta > \theta_0$ pero si cambia para $\theta \neq \theta_0$. La existencia de un test UMP está dada por el teorema de Lehmann, que afirma que existe un test UMP si para un estadístico T el cociente

$$\frac{f(x\,|\theta_1)}{f(x\,|\theta_2)}$$

es una función monótona creciente cuando $\theta_1 > \theta_2$. Esta condición está asegurada con estadísticos suficientes T con una distribución de tipo exponencial.

Prueba de razón de verosimilitudes, que permite extender el caso anterior cuando no existe un test UMP. La prueba de razón de verosimilitudes requiere modelos anidados: modelos en los que el más complejo se puede transformar en el modelo más simple al imponer un conjunto de restricciones a los parámetros. Si los modelos no están anidados, entonces, en general, se puede usar una generalización de la prueba de razón de verosimilitud: la probabilidad relativa. La prueba de razón de verosimilitudes tiene la propiedad deseable de lograr automáticamente la reducción de los datos por suficiencia. La estadística de prueba depende de una estadística mínima suficiente solamente. Esto es inmediato debido a su definición como un cociente y la caracterización de la suficiencia por el teorema de la factorización. Otra propiedad de la prueba es también inmediata: la estadística de

razón de verosimilitudes es invariante bajo las transformaciones del espacio de parámetros que dejan sin cambio la hipótesis nula y la hipótesis alternativa. Este requisito a menudo se impone a las estadísticas de prueba, pero no es necesariamente deseable [6].

Sea la hipótesis nula $H_0 : \theta \in \Theta_0$ contra la hipótesis alternativa $H_1 : \theta \in \Theta \setminus \Theta_0$. Se define la razón de verosimilitudes:

$$\lambda_n = \frac{L_n\left(\theta_n^*\right)}{L_n\left(\widehat{\theta}_n\right)},$$

donde $L_n\left(\theta_n^*\right) = \sup_{\theta\in\Theta_0} \prod_{j=1}^{n} f\left(x_j\,|\theta\right)$, $L_n\left(\widehat{\theta}_n\right) = \sup_{\theta\in\Theta} \prod_{j=1}^{n} f\left(x_j\,|\theta\right)$. Tenemos la propiedad de que $\lambda_n \in [0,1]$ (véase [9]).

El numerador corresponde a la probabilidad de un resultado observado bajo la hipótesis nula. El denominador corresponde a la máxima probabilidad de que un resultado observado varíe los parámetros en todo el espacio de parámetros. El numerador de esta relación es menor que el denominador. Por lo tanto, la razón de verosimilitudes está entre 0 y 1. Los valores bajos de la razón de verosimilitud significan que el resultado observado era menos probable que ocurriera bajo la hipótesis nula en comparación con la alternativa. Los valores altos de la estadística significan que el resultado observado es casi tan probable que ocurra bajo la hipótesis nula como la alternativa, y la hipótesis nula no se puede rechazar. Si la distribución de la razón de verosimilitudes correspondiente a una hipótesis nula y alternativa particular puede determinarse explícitamente, entonces puede usarse directamente para formar regiones de decisión (para aceptar o rechazar la hipótesis nula). Sin embargo, en la mayoría de los casos, la distribución exacta de la razón de verosimilitudes correspondiente a hipótesis específicas es muy difícil de determinar.

2.3. Matriz hessiana. Sea f una función dependiente de un vector $\theta \in \mathbb{R}^k$ de componentes θ^i. Se define Hessiano H_f como la matriz simétrica de las segundas derivadas de la función f con respecto a θ (véase [10])

$$H_f = \left(\frac{\partial^2 f}{\partial\theta^i\partial\theta^j}\right)_{i,j} = \begin{pmatrix} \frac{\partial^2 f}{(\partial\theta^1)^2} & \frac{\partial^2 f}{\partial\theta^1\partial\theta^2} & \cdots & \frac{\partial^2 f}{\partial\theta^1\partial\theta^k} \\ \frac{\partial^2 f}{\partial\theta^2\partial\theta^1} & \frac{\partial^2 f}{(\partial\theta^2)^2} & \cdots & \frac{\partial^2 f}{\partial\theta^2\partial\theta^k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f}{\partial\theta^k\partial\theta^1} & \frac{\partial^2 f}{\partial\theta^k\partial\theta^2} & \cdots & \frac{\partial^2 f}{\left(\partial\theta^k\right)^2} \end{pmatrix}.$$

2.4. Matriz de varianza-covarianza de un vector aleatorio. Sea X un vector real aleatorio de dimensión n, media μ y matriz de varianza-covarianza Σ:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}, \qquad \mathbb{E}(X) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_n \end{pmatrix} = \mu,$$

$$\Sigma = Var(X) = (\sigma_{ij})_{ij} = \mathbb{E}\left((X-\mu)(X-\mu)^T\right) = \mathbb{E}\left(XX^T\right) - \mu\mu^T,$$

donde

$$\sigma_{ij} = \begin{cases} \operatorname{cov}(x_i, x_j) \text{ si } i \neq j, \\ \operatorname{var}(x_i) \text{ si } i = j. \end{cases}$$

La matriz Σ es semi-definida positiva.

2.5. Cantidad de información de Fisher. Sea una variable aleatoria X con función de densidad o probabilidad conjunta $f(x|\theta)$ en donde θ es un parámetro desconocido en el conjunto Θ.

Definición. Se llama cantidad de información de Fisher dada por X sobre el parámetro θ a la cantidad

$$J(\theta) = \mathbb{E}_\theta\left(\Psi(X,\theta)\Psi(X,\theta)^T\right) \text{ es una matriz } k \times k,$$

donde

$$\Psi(X,\theta) = \left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^T \text{ es vector de dimensión } k.$$

También se puede definir la cantidad de información de Fisher $J(\theta)$ de otra manera, la cual se va a usar más adelante (véase [3]).

Lema 2. *Supongamos que las segundas derivadas con respecto a θ de la función $f(x|\theta)$ existen y son continuas. Ademas supongamos que se puede intercambiar los símbolos de la integral y de las derivadas en la integral $\int f(x|\theta)d\nu(x)$. Entonces es válida la siguiente fórmula*

$$J(\theta) = -\mathbb{E}_\theta\left(\dot{\Psi}(X,\theta)\right),$$

donde

$$\dot{\Psi}(X,\theta) = \left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right) \text{ es una matriz } k \times k.$$

Demostración. Por las condiciones de lema tenemos

$$\begin{aligned} \mathbb{E}_\theta[\Psi(X,\theta)] &= \int\left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right)f(x|\theta)d\nu(x) \\ &= \int\frac{\partial}{\partial\theta}f(x|\theta)d\nu(x) = 0, \end{aligned}$$

por lo tanto $J(\theta)$ es matriz de covarianza de Ψ

$$J(\theta) = \operatorname{var}_\theta(\Psi(X,\theta)).$$

Ya que se puede intercambiar las segundas derivadas con respecto al parámetro θ con el signo de la integral, entonces

$$\int \frac{\partial^2}{\partial\theta^2} f(x\,|\theta)d\nu\,(x) = 0,$$

de donde

$$\begin{aligned}
\mathbb{E}_\theta\left(\dot{\Psi}\,(X,\theta)\right) &= \int \left(\frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta} f(x\,|\theta)}{f(x\,|\theta)}\right)\right) f(x\,|\theta)d\nu\,(x) \\
&= \int \left(\frac{\frac{\partial^2}{\partial\theta^2} f(x\,|\theta)}{f(x\,|\theta)}\right) f(x\,|\theta)d\nu\,(x) \\
&- \int \left(\frac{\left(\frac{\partial}{\partial\theta} f(x\,|\theta)\right)^T \left(\frac{\partial}{\partial\theta} f(x\,|\theta)\right)}{\left(f(x\,|\theta)\right)^2}\right) f(x\,|\theta)d\nu\,(x) \\
&= \int \frac{\partial^2}{\partial\theta^2} f(x\,|\theta)d\nu\,(x) \\
&- \int \left(\frac{\partial}{\partial\theta}\log f(x\,|\theta)\right)^T \left(\frac{\partial}{\partial\theta}\log f(x\,|\theta)\right) f(x\,|\theta)d\nu\,(x) \\
&= -\int \left(\Psi\,(X,\theta)\right)\left(\Psi\,(X,\theta)\right)^T f(x\,|\theta)d\nu\,(x) \\
&= -\mathbb{E}_\theta\left(\Psi\,(X,\theta)\,\Psi\,(X,\theta)^T\right) = -J\,(\theta)\,.
\end{aligned}$$

Q.E.D. □

2.6. Distribución normal multivariada $\mathbf{N}_n(\mathbf{0};\mathbf{\Sigma}_n)$. Se supone ahora que el vector $X \in \mathbb{R}^n$ es un vector normal. Se puede definir la distribución normal multivariada (véase [4]).

Definición 3. *Se dice que X tiene distribución normal multivariada $\mathbf{N}_n(0,\Sigma_n)$, donde Σ_n es una matriz $n \times n$, si y sólo si su función característica es:*

$$\varphi_X\,(t) = \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma_n t\right).$$

2.7. La distribución χ^2_n. Si X tiene distribución normal $\mathbf{N}_n(\mathbf{0},\mathbf{I}_n)$, donde I_n es matriz unitaria $n \times n$, entonces (véase [8])

Definición 4. *Se dice que $\|X\|^2 = \sum_{i=1}^n X_i^2$ tiene la distribución χ^2_n.*

3. Preliminares

3.1. Teorema central del límite.

Teorema 5. *([5])Sea un vector aleatorio $(X_1, X_2, ..., X_n) \in \mathbb{R}^n$ de variables i.i.d. con esperanza μ y matriz de covarianza finita Σ_n. Entonces*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N_n\,(0,\Sigma_n)\,,$$

donde $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$ y $N_n\,(0,\Sigma_n)$ es distribución normal.

Demostración. Ya que

$$\sqrt{n}\left(\overline{X}_n-\mu\right)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(X_j-\mu\right),$$

entonces la función característica se puede escribir en la siguente forma

$$\begin{aligned}\varphi_{\sqrt{n}\left(\overline{X}_n-\mu\right)}\left(t\right)&=\varphi_{\sum_{j=1}^{n}\left(X_j-\mu\right)}\left(\frac{t}{\sqrt{n}}\right)\\&=\prod_{j=1}^{n}\varphi_{X_j-\mu}\left(\frac{t}{\sqrt{n}}\right)=\left(\varphi\left(\frac{t}{\sqrt{n}}\right)\right)^n,\end{aligned}$$

donde $\varphi\left(t\right)$ es la función característica de $X_j-\mu$. Como $\varphi\left(0\right)=1$, $\dot{\varphi}\left(0\right)=1$ y

$$\lim_{\varepsilon\to 0}\ddot{\varphi}\left(\varepsilon\right)=-\Sigma_n,$$

usando teorema de Taylor obtenemos

$$\begin{aligned}\varphi_{\sqrt{n}\left(\overline{X}_n-\mu\right)}\left(t\right)&=\left(1+\frac{1}{n}t^T\int_0^1\int_0^1 v\ddot{\varphi}\left(\frac{uvt}{\sqrt{n}}\right)dudv\,t\right)^n\\&\to\exp\left(\lim_{n\to\infty}t^T\int_0^1\int_0^1 v\ddot{\varphi}\left(\frac{uvt}{\sqrt{n}}\right)dudv\,t\right)\\&=\exp\left(-\frac{1}{2}t^T\Sigma_n t\right).\end{aligned}$$

Q.E.D. □

3.2. Ley fuerte uniforme de los números grandes. Sea un vector aleatorio $(X_1,X_2,...,X_n)\in\mathbb{R}^n$ de variables i.i.d. con función de distribución $F(x)$. Sea $U\left(x,\theta\right)$ una función medible de x para todos $\theta\in\Theta$. Suponga que

$$\mu\left(\theta\right)=\mathbb{E}\left[U\left(x,\theta\right)\right]=\int U\left(x,\theta\right)dF\left(x\right)$$

existe y es finito para todos $\theta\in\Theta$. Por la ley fuerte de los números grandes (véase [7]).

$$\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)\overset{c.s.}{\to}\mu\left(\theta\right)$$

cuando $n\to\infty$ para todo $\theta\in\Theta$ fijo. Para nosotros es importante encontrar condiciones que garanticen la convergencia uniforme con respecto al parámetro $\theta\in\Theta$, es decir

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)-\mu\left(\theta\right)\right|\overset{c.s.}{\to}0,$$

cuando $n\to\infty$. El resultado siguiente (vease [6]) nos proporciona tales condiciónes suficientes.

Teorema 6. *Supongamos que*

1) Θ *es compacto,*

2) $U(x,\theta)$ *es semicontinua con respecto a* θ *para todos* x*,*

3) Existe la función $K(x)$ *tal que* $\mathbb{E}[K(x)]$ *está acotada y* $U(x,\theta) \leq K(x)$ *para todos* x *y* θ*,*

4) Para todos θ *y para* $\rho > 0$ *suficientemente pequeño, el* $\sup_{|\theta'-\theta|<\rho} U(x,\theta')$ *es medible con respecto a* x*. Entonces*

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}} \sup_{\theta\in\Theta} \frac{1}{n}\sum_{j=1}^{n} U(X_j,\theta) \leq \sup_{\theta\in\Theta} \mu(\theta)\right) = 1.$$

Demostración. Denotamos

$$\phi(x,\theta,\rho) = \sup_{|\theta'-\theta|<\rho} U\left(x,\theta'\right).$$

Por condición (4) la función ϕ es medible con respecto a x para todo $\rho > 0$ suficientemente pequeño y por condición (3) está acotada por una función integrable. Además $\phi(x,\theta,\rho) \searrow U(x,\theta)$ cuando $\rho \searrow 0$ por 2).

Por lo tanto, por el teorema de la convergencia monótona para $\rho \searrow 0$ obtenemos

$$\int \phi(x,\theta,\rho)\, dF(x) \searrow \int U(x,\theta)\, dF(x) = \mu(\theta).$$

Ahora fijamos $\varepsilon > 0$. Para cada θ se puede encontrar ρ_θ tal que

$$\int \phi(x,\theta,\rho_\theta)\, dF(x) < \mu(\theta) + \varepsilon.$$

Las esferas $S(\theta,\rho_\theta) = \{\theta' : |\theta'-\theta| < \rho_\theta\}$ cubren a Θ. Entonces por la propiedad 1) existe la subcubierta finita, digamos $\Theta \subset \cup_{j=1}^{m} S\left(\theta_j,\rho_{\theta_j}\right)$. Para cada $\theta \in \Theta$ existe un índice j tal que $\theta \in S\left(\theta_j,\rho_{\theta_j}\right)$. Por las condiciones para ϕ, $U(x,\theta) \leq \phi\left(x,\theta_j,\rho_{\theta_j}\right)$ para todos x. Entonces

$$\frac{1}{n}\sum_{j=1}^{n} U(X_j,\theta) \leq \frac{1}{n}\sum_{j=1}^{n} \phi\left(X_j,\theta_j,\rho_{\theta_j}\right),$$

por consiguiente

$$\sup_{\theta\in\Theta} \frac{1}{n}\sum_{j=1}^{n} U(X_j,\theta) \leq \sup_{1\leq j\leq m} \frac{1}{n}\sum_{j=1}^{n} \phi\left(X_j,\theta_j,\rho_{\theta_j}\right).$$

Ahora aplicamos la ley fuerte de los números grandes a $\frac{1}{n}\sum_{j=1}^{n} \phi\left(X_j,\theta_j,\rho_{\theta_j}\right)$. obtenemos

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \phi\left(X_j,\theta_j,\rho_{\theta_j}\right) \leq \mu(\theta_j) + \varepsilon, \text{ para } j = 1,2,...,m\right) = 1,$$

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}}\sup_{1\leq j\leq m}\frac{1}{n}\sum_{j=1}^{n}\phi\left(X_j,\theta_j,\rho_{\theta_j}\right)\leq\sup_{1\leq j\leq m}\mu\left(\theta_j\right)+\varepsilon\right)=1,$$

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}}\sup_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)\leq\sup_{\theta\in\Theta}\mu\left(\theta\right)+\varepsilon\right)=1.$$

Como $\varepsilon>0$ es arbitrario, entonces

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}}\sup_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)\leq\sup_{\theta\in\Theta}\mu\left(\theta\right)\right)=1.$$

Q.E.D. □

Teorema 7. *Supongamos que*
1) Θ *es compacto,*
2) $U\left(x,\theta\right)$ *es continua con respecto a* θ *para todos* x*,*
3) Existe la función $K\left(x\right)$ *tal que* $\mathbb{E}\left[K\left(x\right)\right]$ *está acotada y* $\left|U\left(x,\theta\right)\right|\leq K\left(x\right)$ *para todos* x *y* θ*. Entonces*

$$\mathbb{P}\left(\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)-\mu\left(\theta\right)\right|=0\right)=1.$$

Demostración. Primero notamos que como $U\left(x,\theta\right)$ es continua con respecto a θ, entonces la condición 4) del Teorema 6 es válida, en realidad

$$\sup_{|\theta'-\theta|<\rho}U\left(x,\theta'\right)=\sup_{\theta'\in D}U\left(x,\theta'\right)$$

para cualquier conjunto D numerable y denso en $\left\{\theta':\left|\theta'-\theta\right|<\rho\right\}$. Notamos que $\mu\left(\theta\right)$ es continua, ya que

$$\begin{aligned}\lim_{\theta'\to\theta}\mu\left(\theta'\right)&=\lim_{\theta'\to\theta}\int U\left(x,\theta'\right)dF\left(x\right)\\&=\int\lim_{\theta'\to\theta}U\left(x,\theta'\right)dF\left(x\right)=\int U\left(x,\theta\right)dF\left(x\right)=\mu\left(\theta\right).\end{aligned}$$

Por el teorema de Lebesgue de la convergencia dominada, como U está acotada por K, que es integrable (vease condición 3)). Por lo tanto si el Teorema 6 es válido para $\mu\left(\theta\right)=0$, entonces es válido también para cualquier $\mu\left(\theta\right)$ considerando la diferencia $U\left(x,\theta\right)-\mu\left(\theta\right)$, que es continua con respecto a θ y acotada por la función $K\left(x\right)+\mathbb{E}\left[K\left(x\right)\right]$. Por lo tanto consideramos el caso de $\mu\left(\theta\right)=0$. Por el Teorema 6 aplicado a $U\left(x,\theta\right)$ y $-U\left(x,\theta\right)$ tenemos

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}}\sup_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)\leq 0\right)=1$$

y

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}}\sup_{\theta\in\Theta}-\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)\leq 0\right)=1.$$

De donde obtenemos el resultado del teorema

$$\mathbb{P}\left(\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}U\left(X_j,\theta\right)-\mu\left(\theta\right)\right|=0\right)=1,$$

porque para una función cualquiera $g\left(\theta\right)$, tenemos

$$0\leq\sup_{\theta}\left|g\left(\theta\right)\right|=\text{máx}\left\{\sup_{\theta}g\left(\theta\right),\sup_{\theta}-g\left(\theta\right)\right\}.$$

Q.E.D. □

Lema 8. *Supongamos que se cumplen las condiciónes 1)-5). Entonces*

$$I_n\left(\widehat{\theta}_n\right)\overset{c.s.}{\to}J\left(\theta_0\right)$$

cuando $n\to\infty$.

Demostración. Para demostrar que $I_n\left(\widehat{\theta}_n\right)\overset{c.s.}{\to}J\left(\theta_0\right)$ primero notamos que

$$J\left(\theta_0\right)=-\mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x,\theta_0\right)\right)$$

es continua con respecto a θ por condición 3). Entonces existe $\rho>0$ tal que si $\left|\widehat{\theta}_n-\theta_0\right|<\rho$ entonces

$$\left|J\left(\theta_0\right)+\mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x,\widehat{\theta}_n\right)\right)\right|<\varepsilon.$$

Por el teorema de la ley fuerte de los números grandes (Teorema 7), con probabilidad 1 existe un número N entero tal que para todos $n>N$

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\dot{\Psi}\left(x_j,\widehat{\theta}_n\right)-\mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x,\widehat{\theta}_n\right)\right)\right|<\varepsilon.$$

Ahora suponiendo que N es suficientemente grande tal que para todos $n > N$ $\left|\widehat{\theta}_n - \theta_0\right| < \rho$. Entonces para todos $n > N$

$$\begin{aligned}
\left|I_n\left(\widehat{\theta}_n\right) - J\left(\theta_0\right)\right| &\leq \left| -\frac{1}{n}\sum_{j=1}^{n} \dot{\Psi}\left(x_j, \widehat{\theta}_n\right) \right. \\
&\left. + \mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x, \widehat{\theta}_n\right)\right) - J\left(\theta_0\right) - \mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x, \widehat{\theta}_n\right)\right)\right| \\
&\leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n} \dot{\Psi}\left(x_j, \widehat{\theta}_n\right) - \mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x, \widehat{\theta}_n\right)\right)\right| \\
&+ \left|J\left(\theta_0\right) + \mathbb{E}_{\theta_0}\left(\dot{\Psi}\left(x, \widehat{\theta}_n\right)\right)\right| \\
&\leq 2\varepsilon.
\end{aligned}$$

Por lo tanto, $I_n\left(\widehat{\theta}_n\right) \overset{c.s.}{\to} J\left(\theta_0\right)$, cuando $n \to \infty$. Q.E.D. □

4. Demostracion del Teorema de Wilks

Denote $l_n = \log L_n$. Tomando logaritmo de la fórmula (1.1) obtenemos

$$-2\log\lambda_n = 2\left(l_n\left(\widehat{\theta}_n\right) - l_n\left(\theta_n^*\right)\right).$$

El desarrollo en serie de Taylor de $l_n\left(\theta_n^*\right)$ permite escribir

$$\begin{aligned}
l_n\left(\theta_n^*\right) = l_n\left(\widehat{\theta}_n\right) &+ \dot{l}_n\left(\widehat{\theta}_n\right)\left(\theta_n^* - \widehat{\theta}_n\right) \\
&- \frac{1}{2}n\left(\theta_n^* - \widehat{\theta}_n\right)^T I_n\left(\widehat{\theta}_n\right)\left(\theta_n^* - \widehat{\theta}_n\right) + ...,
\end{aligned}$$

donde $I_n\left(\widehat{\theta}_n\right)$ es la matriz hessiana, definida por

$$I_n\left(\widehat{\theta}_n\right) = -\frac{1}{n}\left(\frac{\partial^2}{\partial\theta^i\partial\theta^j} l_n\left(\widehat{\theta}_n\right)\right).$$

Por Lema 8 tenemos que

$$I_n\left(\widehat{\theta}_n\right) \overset{c.s.}{\to} J\left(\theta_0\right), \tag{4.1}$$

donde $J\left(\theta_0\right)$ es matriz de la informacion de Fisher. Notamos que $\dot{l}_n\left(\widehat{\theta}_n\right) = 0$, dado que $\widehat{\theta}_n$ es el estimador de MV de θ, pero $\dot{l}_n\left(\theta_n^*\right)$ no necesariamente se anula. Ahora

$$\begin{aligned}
l_n\left(\theta_n^*\right) - l_n\left(\widehat{\theta}_n\right) &= -\frac{1}{2}n\left(\theta_n^* - \widehat{\theta}_n\right)^T I_n\left(\widehat{\theta}_n\right)\left(\theta_n^* - \widehat{\theta}_n\right) + ... \\
&\approx -\frac{1}{2}n\left(\theta_n^* - \widehat{\theta}_n\right)^T I_n\left(\widehat{\theta}_n\right)\left(\theta_n^* - \widehat{\theta}_n\right) \\
&\approx -\frac{1}{2}n\left(\theta_n^* - \widehat{\theta}_n\right)^T J\left(\theta_0\right)\left(\theta_n^* - \widehat{\theta}_n\right).
\end{aligned}$$

Por lo tanto

$$-2\log\lambda_n = 2\left(l_n\left(\widehat{\theta}_n\right) - l_n\left(\theta_n^*\right)\right) = n\left(\theta_n^* - \widehat{\theta}_n\right)^T J\left(\theta_0\right)\left(\theta_n^* - \widehat{\theta}_n\right).$$

Ahora buscamos la distribución asintótica de $\sqrt{n}\left(\theta_n^* - \widehat{\theta}_n\right)$. Expandiendo $\dot{l}_n\left(\theta_n^*\right)$ en serie de Taylor alrededor de $\widehat{\theta}_n$ obtenemos

$$\dot{l}_n\left(\theta_n^*\right) \approx \dot{l}_n\left(\widehat{\theta}_n\right) + \ddot{l}_n\left(\widehat{\theta}_n\right)\left(\theta_n^* - \widehat{\theta}_n\right) = \ddot{l}_n\left(\widehat{\theta}_n\right)\left(\theta_n^* - \widehat{\theta}_n\right). \tag{4.2}$$

Por (4.1) se puede escribir

$$-\frac{1}{n}\ddot{l}_n\left(\widehat{\theta}_n\right) = -\frac{1}{n}\left(\frac{\partial^2}{\partial\theta^i\partial\theta^j}l_n\left(\widehat{\theta}_n\right)\right) = I_n\left(\widehat{\theta}_n\right) \overset{c.s.}{\to} J\left(\theta_0\right).$$

De la fórmula (4.2) obtenemos

$$\theta_n^* - \widehat{\theta}_n = \left(\ddot{l}_n\left(\widehat{\theta}_n\right)\right)^{-1}\dot{l}_n\left(\theta_n^*\right),$$

por lo tanto

$$\begin{aligned}
\sqrt{n}\left(\theta_n^* - \widehat{\theta}_n\right) &= \sqrt{n}\left(\ddot{l}_n\left(\widehat{\theta}_n\right)\right)^{-1}\dot{l}_n\left(\theta_n^*\right) \\
&= \left(-\frac{1}{n}\ddot{l}_n\left(\widehat{\theta}_n\right)\right)^{-1}\left(-\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right) \\
&= -\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right).
\end{aligned}$$

Entonces

$$\begin{aligned}
-2\log\lambda_n &= n\left(\theta_n^* - \widehat{\theta}_n\right)^T J\left(\theta_0\right)\left(\theta_n^* - \widehat{\theta}_n\right) \\
&= \left(-\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right)^T J\left(\theta_0\right)\left(-\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right) \\
&= \left(\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right)^T J\left(\theta_0\right)\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right) \\
&= \left(\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right)^T \frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right).
\end{aligned}$$

Notamos que

$$\begin{aligned}
\left(\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right)^T &= \left(\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right)^T\left(\left(J\left(\theta_0\right)\right)^{-1}\right)^T \\
&= \left(\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)\right)^T\left(J\left(\theta_0\right)\right)^{-1},
\end{aligned}$$

ya que el hessiano es simétrico bajo la condición 2). Por lo tanto, obtenemos

$$-2\log\lambda_n = \frac{1}{\sqrt{n}}\left(\dot{l}_n\left(\theta_n^*\right)\right)^T\left(J\left(\theta_0\right)\right)^{-1}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right), \tag{4.3}$$

donde el vector $\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)$ tiene la siguiente forma

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta^1}\log f\left(x_i\left|\theta_n^*\right.\right),...,\sum_{i=1}^{n}\frac{\partial}{\partial\theta^k}\log f\left(x_i\left|\theta_n^*\right.\right)\right).$$

Vamos a buscar la distribución asintótica de $\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)$ alrededor de θ_0. Expandimos en serie de Taylor con el centro en θ_0

$$\begin{aligned}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right) &\approx \frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_0\right)+\frac{1}{\sqrt{n}}\ddot{l}_n\left(\theta_0\right)\left(\theta_n^*-\theta_0\right)\\ &= \frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_0\right)-J\left(\theta_0\right)\sqrt{n}\left(\theta_n^*-\theta_0\right). \qquad (4.4)\end{aligned}$$

Denotamos las matrices:

$$G_1 \text{ de dimensión } r\times r,$$

$$G_2 \text{ de dimensión } r\times(k-r),$$

$$G_3 \text{ de dimensión } (k-r)\times(k-r),$$

tales que la matriz

$$J\left(\theta_0\right)=-\mathbb{E}_{\theta_0}\left(\frac{\partial^2}{\partial\theta^2}\log f\left(x\left|\theta_0\right.\right)\right)$$

de dimension $k\times k$ se puede escribir en la siguiente forma

$$J\left(\theta_0\right)=\begin{pmatrix} G_1 & G_2\\ G_2^T & G_3\end{pmatrix}.$$

Definimos la matriz $\widetilde{H}$ (análogamente a la matriz $J\left(\theta_0\right)$)

$$\widetilde{H}=\begin{pmatrix} 0 & 0\\ 0 & G_3^{-1}\end{pmatrix}.$$

Puesto que los últimos $k-r$ componentes de $\dot{l}_n\left(\theta_n^*\right)$ son ceros, entonces

$$\widetilde{H}\dot{l}_n\left(\theta_n^*\right)=0.$$

Por lo tanto de la ecuación (4.4) obtenemos

$$\begin{aligned}\widetilde{H}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_0\right) &= \widetilde{H}\frac{1}{\sqrt{n}}\dot{l}_n\left(\theta_n^*\right)+\widetilde{H}J\left(\theta_0\right)\sqrt{n}\left(\theta_n^*-\theta_0\right)\\ &= \widetilde{H}J\left(\theta_0\right)\sqrt{n}\left(\theta_n^*-\theta_0\right).\end{aligned}$$

Puesto que las primeras r componentes de θ_n^* y θ_0 son nulas, entonces

$$\begin{aligned}
&\widetilde{H} J(\theta_0) \sqrt{n}(\theta_n^* - \theta_0) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & G_3^{-1} \end{pmatrix} \begin{pmatrix} G_1 & G_2 \\ G_2^T & G_3 \end{pmatrix} \sqrt{n}(\theta_n^* - \theta_0) \\
&= \begin{pmatrix} 0 & 0 \\ G_3^{-1} G_2^T & G_3^{-1} G_3 \end{pmatrix} \sqrt{n}(\theta_n^* - \theta_0) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & G_3^{-1} G_3 \end{pmatrix} \sqrt{n}(\theta_n^* - \theta_0) \\
&= \sqrt{n}(\theta_n^* - \theta_0),
\end{aligned}$$

es decir

$$\widetilde{H} \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = \sqrt{n}(\theta_n^* - \theta_0).$$

Sustituyendo la última igualdad en (4.4), encontramos

$$\begin{aligned}
\frac{1}{\sqrt{n}} \dot{l}_n(\theta_n^*) &= \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) - J(\theta_0) \sqrt{n}(\theta_n^* - \theta_0) \\
&= \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) - J(\theta_0) \widetilde{H} \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) \\
&= \frac{1}{\sqrt{n}} \left(I - J(\theta_0) \widetilde{H} \right) \dot{l}_n(\theta_0).
\end{aligned}$$

Por teorema central del límite (Teorema 5)

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = \sqrt{n} \left(\frac{1}{n} \dot{l}_n(\theta_0) \right) \xrightarrow{d} N_k(0, J(\theta_0)).$$

Entonces

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_n^*) \xrightarrow{d} \left(I - J(\theta_0) \widetilde{H} \right) Y,$$

donde la variable aleatoria Y tiene distribución normal $\mathbf{N}_k(0, J(\theta_0))$. De la ecuación (4.3) tenemos que

$$-2 \log \lambda_n \xrightarrow{d} Y^T \left(I - J(\theta_0) \widetilde{H} \right)^T (J(\theta_0))^{-1} \left(I - J(\theta_0) \widetilde{H} \right) Y.$$

Puesto que

$$\begin{aligned}
&\left(I - J(\theta_0) \widetilde{H} \right)^T (J(\theta_0))^{-1} \left(I - J(\theta_0) \widetilde{H} \right) \\
&= \left(I - \widetilde{H}^T J(\theta_0) \right) (J(\theta_0))^{-1} \left(I - J(\theta_0) \widetilde{H} \right) \\
&= \left((J(\theta_0))^{-1} - \widetilde{H}^T \right) \left(I - J(\theta_0) \widetilde{H} \right) \\
&= (J(\theta_0))^{-1} - \widetilde{H} - \widetilde{H}^T + \widetilde{H}^T J(\theta_0) \widetilde{H} \\
&= (J(\theta_0))^{-1} - \widetilde{H},
\end{aligned}$$

ya que $\widetilde{H}^T J(\theta_0)\widetilde{H} = \widetilde{H} = \widetilde{H}^T$. Entonces,

$$\begin{aligned} &-2\log\lambda_n \xrightarrow{d} Y^T\left(\left(J(\theta_0)\right)^{-1} - \widetilde{H}\right)Y \\ &= Z^T\left(J(\theta_0)\right)^{\frac{1}{2}}\left(\left(J(\theta_0)\right)^{-1} - \widetilde{H}\right)\left(J(\theta_0)\right)^{\frac{1}{2}}Z, \end{aligned}$$

donde $Z = \left(J(\theta_0)\right)^{-\frac{1}{2}}Y$ tiene distribución normal $N_k(0, I)$. La matriz

$$P = \left(J(\theta_0)\right)^{\frac{1}{2}}\left(\left(J(\theta_0)\right)^{-1} - \widetilde{H}\right)\left(J(\theta_0)\right)^{\frac{1}{2}}$$

es idempotente y su rango es igual a

$$\begin{aligned} \text{rango}(P) &= \text{trasa}\left(J(\theta_0)\left(\left(J(\theta_0)\right)^{-1} - \widetilde{H}\right)\right) \\ &= \text{trasa}\left(\left(I - J(\theta_0)\widetilde{H}\right)\right) = r. \end{aligned}$$

Por lo tanto

$$-2\log\lambda_n \xrightarrow{d} Z^T P Z$$

donde $Z^T P Z$ tiene distribución χ^2_r. Q.E.D.

Referencias

[1] Casella, G.; Berger, R.L. (2001). *Statistical inference*. Cengage Learning.

[2] Chernoff, H. (1954). *On the distribution of the likelihood ratio statistic*. Ann. Math. Statist. 25, 573-578.

[3] Cox, D.R.; Hinkley, D.V. (1974). *Theoretical statistics*. Chapman and Hall.

[4] Dudley, R.M. (2003). *Mathematical statistics*. Lecture notes. MIT Open Course Ware.

[5] Feller, W. (1968). *An introduction to probability theory and its applications*, Wiley.

[6] Ferguson, T.S. (1996). *A course in large sample theory*. Chapman & Hall.

[7] Le Cam, L.; Yang, G. (2000). *Asymptotics in Statistics: Some basic concepts*. Springer.

[8] Mood, A.M.; Graybill, F.A. (1963). *Introduction to the theory of statistics*. McGraw-Hill.

[9] Prakasa Rao, B.L.S. (1987). *Asymptotic theory of statistical inference*. Wiley.

[10] Sen, P.K.; Singer, J.M. (1994). *Large sample methods in statistics: An introduction with applications*. Chapman and Hall.

[11] van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

[12] Wilks, S.S. (1938). *The large-sample distribution of the likelihood ratio for testing composite hypotheses*. Ann. Math. Statist. 9, 60-62.

[13] Wilks, S.S. (1962). *Mathematical statistics*. Wiley.