ESTAD_T23. MÉTODO de ESTIM. de MÁXIMA VEROSIMILITUD. PROPIEDADES. DISTRIBUCIÓN ASINTÓTICA de los EMV.

1_ INTRODUCCIÓN

Inferencia →

Estimación →

Contrastación →

Estimac. puntual →

Estim. por intervalos →

Las propiedades de los estimadores son indicadores de la bondad del estimador, pero antes hace falta disponer de métodos objetivos de obtención de estimadores.

El mt. de máximo verosimilitud es uno de ellos.

0. INFERENCIA. INTRODUCCIÓN

~~1. ESTIMACIÓN PUNTUAL I.~~

Recordemos brevemente que la Inferencia Estadística consiste en ~~generalizar~~ (sacar conclusiones) sobre la poblac. que nos interesa estudiar a partir ~~de~~ la información que nos proporciona una muestra aleatoria basándonos en la Teoría de la Probabilidad.

Si estamos interesados en ~~en~~ estudiar el valor de una característica poblacional θ, la Inferencia que podemos utilizar es:
- Estimación,
- Contrastación

La Estimación consiste en dar un valor aproximado del parám. poblacional a partir de la información ~~proporcionada~~ muestral.
La Contrastación consiste en formular una conjetura (hipótesis) sobre el valor del parámetro poblacional y utilizar la inform. muestral para aceptar ó rechazar dicha hipótesis.

En el caso de la estimación, se puede aproximar:
- Estim. puntual → valor concreto
- Estim. por intervalos de confianza → intervalo

En ambos métodos se utiliza un estadístico (función real de la muestra) para estimar.
La ventaja de la estimación puntual es que da un valor concreto del estimador que puede sustituirse directamente en el parámetro, pero sólo cuenta con las propiedades como garantía.
La ventaja de la estimación por intervalos es que el intervalo va acompañado de un grado de confianza de que el verdadero valor del parámetro se encuentre dentro del intervalo, pero su inconveniente es que ofrece infinitas soluciones.

2. CONCEPTO de VEROSIMILITUD

~~[illegible]~~

Sea ξ, una v.a. (discreta/continua) que modeliza el comportamiento de una población ~~[illegible]~~ a partir de una función de distribución $F(x,\theta)$ y de una función de probabilidad (Función de cuantía ó Función de densidad), que contiene un parámetro, o varios, de valor desconocido.

La información disponible para estimar el parámetro es proporcionada por una m.a.s.(n) $X = \{X_1 \dots X_n\}$.

Definimos:

Probabilidad conjunta de la muestra: $P(X/\theta_0)$, es la probab. de extraer la muestra X para un valor del parámetro concreto, θ_0.

$$P(X/\theta) = P(X_1 = x_1, X_2 = x_2 \dots, X_n = x_n / \theta = \theta_0)$$

Para un ~~[illegible]~~ valor fijo del parámetro θ dentro del espacio paramétrico, calculamos la probabilidad de extraer cada una de las muestras posibles.

θ fijo, X variable.

Función de verosimilitud conjunta de la muestra, $\mathcal{L}(\theta/X)$ expresa la verosimilitud, plausibilidad, de que el parámetro presente un valor concreto tomando como base la informac. contenida en la muestra.

$$\mathcal{L}(X,\theta) = \mathcal{L}(x_1 \dots x_n, \theta) = \mathcal{L}(\theta/X).$$

Para una muestra concreta X, calculamos la verosimilitud de que el parámetro tome un valor concreto.

X fija, θ variable.

Es importante distinguir entre probabilidad y verosimilitud. La probabilidad de un suceso se considera como una medida racional de la creencia de que el suceso ocurrirá. Sin embargo, con verosimilitud nos referimos al valor del parámetro que hace más creíble o verosímil una muestra dada. La probabilidad de la muestra es 1, ya que ya ha ocurrido, nos planteamos qué valor de θ la hace más verosímil. La verosimilitud mide nuestro orden de preferencia entre diferentes poblaciones.

La verosimilitud se mide calculando las probabilidades de la muestra para cada parámetro dado y se define como proporcional al suceso $\{X/\theta\}$.

$$\mathcal{L}(\theta/x) = k P(X/\theta)$$

Mientras $P(X/\theta)$ es una auténtica probabilidad, cumple los axiomas y las propiedades, $\sum_i P(X_i/\theta) = 1$, es una función de X; por el contrario $\mathcal{L}(\theta/x)$ es una función del parámetro y sumada o integrada en todo el campo de variación de θ no da 1.

El desarrollo matemático del concepto de verosimilitud se rige por dos principios:

1 – Ley de verosimilitud → Una muestra informa mejor ~~que otra~~ sobre un valor del parámetro que sobre otro si

$$\mathcal{L}(\theta_1/x) > \mathcal{L}(\theta_2/x)$$

2 – Principio de verosimilitud → Toda la información contenida en la muestra para elegir entre dos parámetros se encuentra en la razón de verosimilitud:

$$\lambda(X) = \frac{\mathcal{L}(\theta_1/x)}{\mathcal{L}(\theta_2/x)} \rightarrow \begin{cases} \lambda > 1 \rightarrow \theta_1 + \text{verosímil que } \theta_2 \\ \lambda = 1 \rightarrow \theta_1 = \text{verosímil que } \theta_2 \\ \lambda < 1 \rightarrow \theta_1 - \text{verosímil que } \theta_2 \end{cases}$$

?3. MÉTODO de ESTIMACIÓN de MÁXIMA VEROSIMILITUD

Es, desde el punto de vista teórico, el método general de estimación más conocido. Fue utilizado por Gauss en casos particulares, pero como método de estimación fue introducido por Fisher en 1922, aunque son muy importantes las contribuciones posteriores realizadas por otros autores.

El mt. se fundamenta en el supuesto siguiente: de varios sucesos ocurrirá el más probable.

El método de máxima verosimilitud consiste en elegir como estimador del parámetro desconocido θ aquel valor $\hat{\theta}(x_1 \dots x_n)$ que hace máxima la función de verosimilitud $\mathcal{L}(x_1 \dots x_n; \theta)$.

El planteamiento es resolver el problema:

$$\max_{\theta} \mathcal{L}(x_1 \dots x_n, \theta) \longrightarrow \hat{\theta}_{EMV} \text{ solución}$$

Al estimador $\hat{\theta}(x_1 \dots x_n)$ se le llama estimador máximo-verosímil o estimador máxima verosimilitud (EMV) del parámetro θ.

En general, la función de verosimilitud es complicada y, al ser la ~~trans~~ $\mathcal{L}(x,\theta)$ una función positiva y la transformación logarítmica monótona creciente, ambas funciones tomarán máximos en el mismo punto.

$$\max_{\theta} \mathcal{L}(X,\theta) \sim \max_{\theta} \ln \mathcal{L}(X,\theta)$$

Definimos la función de verosimilitud de n variables aleat. como la función de probab / densidad conjunta:

$$\mathcal{L}(\theta/X) = \mathcal{L}(X,\theta) = \mathcal{L}(x_1 \dots x_n, \theta) = \begin{cases} P(\xi_1 = x_1 \dots \xi_n = x_n, \theta) \\ f(x_1 \dots x_n, \theta) \end{cases}$$

Para una m.a.s., al ser indep. las variables

$$\mathcal{L}(x_1 \dots x_n, \theta) = \begin{cases} P(\xi = x_1 \dots \xi = x_n, \theta) = \prod_{i=1}^{n} P(\xi = x_i, \theta) \\ f(x_1 \dots x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta) \end{cases}$$

la transformación logarítmica tiene la ventaja de que transforma productos en sumas. Si la muestra es aleatoria simple la verosimilitud conjunta se obtiene como el producto de probab. individuales, por lo que:

$$\max_{\theta} \ln \mathcal{L}(x_1 \dots x_n; \theta) = \max \ln \prod_{i=1}^{n} f(x_i; \theta) =$$

$$= \max_{\theta} \sum_{i=1}^{n} \ln f(x_i; \theta)$$

Este problema de optimización se resuelve derivando, suponiendo que se cumplen las condiciones de regularidad:

- Campo de variación de θ es un int. abierto del eje real.
- El campo de variación de ξ no depende de θ
- $f(x;\theta)$ es positiva y derivable respecto a θ
- ~~máx~~ $\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\theta<\hat{\theta}} < 0 \longrightarrow$ condic. de máximo. (nf.)

Así, sólo bastará desarrollar la condición ~~necesaria~~ ~~suficiente~~ y necesaria despejar $\hat{\theta}$

CN $\rightarrow \frac{\partial \ln \mathcal{L}(X;\theta)}{\partial \theta} = 0 \iff$ ~~$\theta = \hat{\theta}$ máximo~~ $\theta = \hat{\theta}(x_1 \dots x_n)$ EMV(θ).

$$\frac{\partial \ln \mathcal{L}(X;\theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} \ln f(x_i;\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta} = 0.$$

las soluciones a esta ecuación no tienen por qué ser únicas.

1 soluc $\longrightarrow \hat{\theta}$ EMV en sentido estricto

Varias soluc $\longrightarrow \hat{\theta}_i$ EMV en sentido amplio

$\hat{\theta}$ es una función muestral, por lo que prescindiremos de las soluciones que den lugar a una cte.

ESTAD _ T23

Si la función de probabilidad (densidad/cuantía) de la población depende de K parámetros, entonces los EMV de estos parámetros se obtienen resolviendo el sistema de ecuaciones:

$$\left.\begin{array}{l} \dfrac{\partial \ln \mathcal{L}(X,\theta)}{\partial \theta_1} = 0 \\ \cdots\cdots \\ \dfrac{\partial \ln \mathcal{L}(X,\theta)}{\partial \theta_K} = 0 \end{array}\right\} \xrightarrow{\text{soluc.}} \begin{array}{l} \hat{\theta}_1 = \hat{\theta}_1(x_1 \dots x_n) \\ \cdots\cdots \\ \hat{\theta}_K = \hat{\theta}_K(x_1 \dots x_n) \end{array}$$

Ejemplos:

$\xi \to B(m, p) \Rightarrow \hat{p}_{MV} = \dfrac{\bar{X}}{m}$

$\xi \to \mathcal{P}(\lambda) \Rightarrow \hat{\lambda}_{MV} = \bar{X}$

$\xi \to N(\mu, \sigma)$ σ conocida $\Rightarrow \hat{\mu} = \bar{X}$

$\xi \to N(\mu, \sigma)$ μ conocido $\Rightarrow \hat{\sigma}^2 = \dfrac{\sum(x_i - \mu)^2}{n} \neq S^2$.

$\xi \to N(\mu, \sigma)$ μ y σ descon. $\Rightarrow \begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = S^2 \text{ (varianza muestral)} \end{cases}$ (lo hago al final)

No siempre se puede obtener el ~~estim~~ EMV por el método analítico expuesto.

Ejemplo: $\xi \to \mathcal{U}(0, \theta)$ ¿$\hat{\theta}_{MV}$?

$f(x) = \dfrac{1}{b-a} = \dfrac{1}{\theta} \Rightarrow \mathcal{L}(x_1 \dots x_n, \theta) = \dfrac{1}{\theta^n}$, no depende de X

$\hat{\theta}$ no es admisible

La solución se obtiene partiendo de la idea del método: es más verosímil lo que ya ha ocurrido.

Si la muestra es $X = \{x_1 \dots x_n\}$ y cada x_i / $0 < x_i < \theta$, como ningún x_i puede ser mayor que θ, lo más verosímil es $\hat{\theta}_{MV} = \max\limits_{i=1..n} \{x_i\}$.

4. PROPIEDADES de los ESTIMADORES MÁXIMO VEROSÍMILES

Tal y como se ha definido, el estimador máximo verosím. coincide con la Moda de la función de verosimilitud. Generalmente, la moda es más pobre que la media o que la mediana, por lo que se justifica que los EMV tengan propiedades "pobres" para muestras pequeñas y para muestras grandes tengan muchas más, ya que la moda tiende a la media, si existe.

1- Insesgadez

En general, los EMV NO son insesgados.
Sin embargo, SÍ son ASINT INSESG, como se deduce de la propiedad de consistencia.

$\hat{\theta}$ es $\left| \begin{matrix} \text{insesgado de } \theta \\ \text{asint. insesgado} \end{matrix} \right|$ si $\left| \begin{matrix} E[\hat{\theta}] = \theta \\ E[\hat{\theta}] \xrightarrow[n \to \infty]{} \theta \end{matrix} \right|$

2- Consistencia

Bajo condiciones generales los EMV SÍ son CONSISTENTES.

$$\forall \varepsilon > 0, \quad \lim_{n \to \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1, \quad \forall \theta$$

Consistencia en probabilidad

La demostración no es trivial, se basa en aplicar la desig. de Jensen ($E(\ln) \le \ln(E)$) al cociente de verosimilitudes y en la Ley Fuerte de los Grandes Números.

Si $\hat{\theta}_{MV}$ no es insesgado, al ser consistente será automáticamente asintóticamente insesgado.

3_ Eficiencia

Si existe un estimador eficiente de θ $(V(\hat{\theta}) = CCR)$, entonces coincide con el estim. Máx. Verosimilitud.

Dem: Según las condiciones de regularidad de Wolfowitz, se llega a que la segunda derivada del logaritmo de la función de verosimilitud es <0 en $\theta = \hat{\theta}_{MV}$, porque anula la 1ª derivada.

4_ Normalidad y Eficiencias asintóticas

Los estim. máx-verosímiles SI son asintóticamente normales con esperanza θ y asintóticamente eficientes.

La normalidad la vemos luego, pero

$$V(\hat{\theta}) \xrightarrow[n \to \infty]{} \frac{1}{I(\theta)} \equiv CCR \Rightarrow \text{asintót. eficientes}$$

5_ Suficiencia

Si $T = T(X)$ es un estad. suficiente de θ, entonces $\hat{\theta}_{MV}$ es función de $T(X)$.

Partiendo de la suficiencia de T, descomponemos la func. verosimilitud por teo. factorización, tomando logaritmos, derivando e igualando a 0 se llega a $\hat{\theta}$ función de T.

6_ Invarianza

Si $\hat{\theta}$ es el EMV de θ, entonces $g(\hat{\theta})$ es el EMV de $g(\theta)$.

5. DISTRIBUCIÓN ASINTÓTICA de los EMV

Los estimadores máximo-verosímiles son asintóticamente normales con esperanza θ. Lo vemos:

Por las condiciones de regularidad de Wolfowitz:

$$E\left[\frac{\partial \ln \mathcal{L}(X,\theta)}{\partial\theta}\right]=0$$

y además existe la cantidad de información de Fisher:

$$I(\theta)=-E\left[\frac{\partial^2 \ln \mathcal{L}(X,\theta)}{\partial\theta^2}\right]$$

Si $\hat{\theta}$ es el EMV, desarrollando por serie de Taylor la derivada parcial en un entorno de θ_0, verdadero valor del parámetro:

$$\left.\frac{\partial \ln \mathcal{L}(X,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}=\left.\frac{\partial \ln \mathcal{L}(X,\theta)}{\partial\theta}\right|_{\theta=\theta_0}+(\hat{\theta}-\theta_0)\cdot\left.\frac{\partial^2 \ln \mathcal{L}(X,\theta)}{\partial\theta^2}\right|_{\theta=\theta'}$$

siendo θ' un valor intermedio entre $\hat{\theta}$ y θ_0.

Por ser $\hat{\theta}$ el EMV, hace la primera derivada 0:

$$\left.\frac{\partial \ln \mathcal{L}(X,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}=0$$

que era el 1er miembro de la igualdad. Igualando a 0 y reordenando resulta:

$$\hat{\theta}-\theta_0=-\frac{\left.\frac{\partial \ln \mathcal{L}(X,\theta)}{\partial\theta}\right|_{\theta=\theta_0}}{\left.\frac{\partial^2 \ln \mathcal{L}(X,\theta)}{\partial\theta^2}\right|_{\theta=\theta'}}$$

Multiplicando ~~por~~ $(\sqrt{I}\cdot\sqrt{I})$ y dividiendo por $I(\theta)$:

(*) $$(\hat{\theta}-\theta_0)\sqrt{I(\theta)}=\frac{\left.\frac{\partial \ln \mathcal{L}(X,\theta)}{\partial\theta}\right|_{\theta=\theta_0}\Big/\sqrt{I(\theta)}\cdot\cancel{\sqrt{I(\theta)}}}{\left.\frac{\partial^2 \ln \mathcal{L}(X,\theta)}{\partial\theta^2}\right|_{\theta=\theta'}\Big/-I(\theta)}\qquad \xrightarrow[n\to\infty]{c.s.} 1$$

Por otra parte, recordamos que:

$$E\left[\frac{\partial \ln \mathcal{L}(\bar{X},\theta)}{\partial \theta}\right] = 0 \qquad y \qquad V\left[\frac{\partial \ln \mathcal{L}(\bar{X},\theta)}{\partial \theta}\right] = I(\theta)$$

Si consideramos la función de verosimilitud como producto de funciones de probabilidad,

$$\frac{\partial \ln \mathcal{L}(\bar{X},\theta)}{\partial \theta} = \underbrace{\sum_{i=1}^{n} \frac{\partial \ln f(X_i;\theta)}{\partial \theta}}_{\substack{\text{suma de n v.a.i.i.d.} \\ \text{por ser m.a.s., se} \\ \text{aplica TCL}}}\Bigg|_{\theta=\theta_0} \xrightarrow[n\to\infty]{d} N\left(0, \sqrt{I(\theta_0)}\right)$$

y por consiguiente

$$\frac{\left.\dfrac{\partial \ln \mathcal{L}(\bar{X},\theta)}{\partial \theta}\right|_{\theta=\theta_0}}{\sqrt{I(\theta_0)}} \xrightarrow{d} N(0,1)$$

Volviendo a (*)

$$(\hat{\theta}-\theta_0)\left(\sqrt{I(\theta_0)}\right) \rightarrow \frac{\dfrac{\partial \ln \mathcal{L}(\bar{X},\theta)}{\partial \theta}}{\sqrt{I(\theta_0)}} \xrightarrow{d} N(0,1)$$

lo que indica que $\hat{\theta}_{MV}$ converge a una distrib. normal.

$$\hat{\theta} \xrightarrow{d} N\left(\theta_0, \frac{1}{\sqrt{I(\theta_0)}}\right)$$