Método de máxima verosimilitud

Curso de Estadística

TAE,2005

J.J. Gómez Cadenas

Muestras

Considerar una variable aleatoria x descrita por la pdf f(x).

El espacio de muestras está constituido por todos los posibles valores de x.

Un conjunto de n observaciones independientes de x se llama una muestra de tamaño n.

Es posible definir un nuevo espacio de muestras constituido por todos los posibles valores del vector $\boldsymbol{x} = (x_1,\ldots,x_n)$. Es decir, la muestra se considera formada por una sola medida aleatoria, caracterizada por las cantidades $(x_1,\ldots,x_n)$.

Las n medidas son independientes

La pdf es la misma para cada medida

$$f_{muestra}(x_1,\ldots x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

Estimadores

Considerar la situación donde se han realizado n medidas de una variable aleatoria cuya pdf se desconoce.

El problema central de la estadística es inferir las propiedades de f(x) basándose en las observaciones $x_1,...,x_n$).

Específicamente, deseamos construir funciones de los x_i para estimar las propiedades de f(x).

A menudo se tiene una hipótesis para la pfd $f(x;\theta)$ de un parámetro desconocido (o más generalmente de un vector de parámetros $\theta=(\theta_1,...,\theta_n)$).

El objetivo es entonces construir funciones de los x_i que permitan estimar los parámetros θ.

Una función de $x_1,...,x_n$ que no contiene parámetros desconocidos se denomina estadística.

Una estadística que se utiliza para estimar una propiedad de una pdf (media, varianza, etc.) se llama un estimador.

Notación: El estimador de un parámetro θ (cuyo valor exacto no se conoce ni es obvio que pueda, en general conocerse) se suele notar como $\overline{\theta}$

Decimos que un estimador es consistente si converge al valor auténtico del parámetro en el límite de alto n: (límite de muestra grande o límite asintótico).

$$\lim_{n\to\infty} P\left(\left|\overline{\theta} - \theta\right| > \varepsilon\right) = 0$$

El procedimiento por el cual estimamos el valor de un parámetro $\overline{\theta}$ a partir de los datos $x_1,...,x_n$ se denomina ajuste (de los datos al parámetro).

Puesto que un estimador $\overline{\theta}(x_1,...,x_n)$ es una función de variables aleatorias, en en sí mismo una variable aleatoria. Es decir, si el experimento se remite muchas veces, para cada muestra $x=(x_1,...,x_n)$ el estimador $\overline{\theta}$ tomará valores diferentes, distribuidos de acuerdo a cierta pdf $g(\overline{\theta};\theta)$ que depende del auténtico valor de parámetro. Esta pdf se denomina distribución de muestreo.

Sesgo

El valor esperado de un estimador $\bar{\theta}$ con pdf g($\bar{\theta};\theta$) es:

$$E[\bar{\theta}(\vec{x})] = \int \bar{\theta} g(\bar{\theta};\theta) d\bar{\theta} = \int \cdots \int \bar{\theta}(\vec{x}) f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n$$

Definimos el sesgo del estimador $\bar{\theta}$ como:

$$b = E[\bar{\theta}(\vec{x})] - \theta$$

NB: El sesgo no depende de los valores $x_1,...,x_n$ de la muestra, sino del tamaño de ésta, de la forma funcional del estimador y de la pdf conjunta (que en general no se conoce).

Decimos que un parámetro no tiene sesgo si b=0 independientemente del tamaño de la muestra.

Decimos que un parámetro no tiene sesgo en el límite asintótico si b=0 cuando n tiene a infinito.

Un parámetro consistente pede sin embargo estar sesgado (n finito)

Estimadores para la media: Media aritmética

Media aritmética o muestral:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ley (débil) de los números grandes: Si existe la varianza de x entoces $\bar{x}$ es un estimador consistente de la media poblacional μ.

$$\lim_{n\to\infty} \bar{x} = \mu$$

Valor esperado de $\bar{x}$:

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu$$

Por lo tanto la media muestral $\bar{x}$ es un estimador sin sesgo de la media poblacional μ.

Estimadores para la varianza y covarianza

Varianza muestral:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\left\langle x^2 \right\rangle - \left\langle x \right\rangle^2 \quad \text{donde } \bar{x}=\left\langle x \right\rangle$$

Al igual que para la media, puede demostrarse que la varianza muestral es un estimador sin sesgo de la varianza poblacional σ^2 . Si la media se conoce entonces también es un estimador sin sesgo la cantidad S^2.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 = \left\langle x^2 \right\rangle - \left\langle \mu \right\rangle^2$$

Análogamente, un estimador sin sesgo para la covarianza es:

$$\hat{V}_{xy}\ \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1}\left\langle xy \right\rangle - \left\langle x \right\rangle\left\langle y \right\rangle$$

Varianza de la media

Dado un estimador, su varianza se define como:

$$V[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

Varianza de la media aritmética:

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] - \mu^2$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} E[x_i x_j] - \mu^2 = \frac{1}{n^2}\left[(n^2 - n)\mu^2 + n(\mu^2 + \sigma^2)\right] - \mu^2 = \frac{\sigma^2}{n}$$

$$(NB: \quad E[x_i x_j] = \mu^2 \;\; i \neq j, \quad E[x_i^2] = \mu^2 + \sigma^2)$$

Es decir: la desviación estándar de la media de n medidas de x es igual a la desviación estándar de f(x) dividida por $\sqrt{n}$.

Varianza de s^2

$$V[s^2] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\mu_2^3)$$

$$(\mu_n = \int_{-\infty}^{+\infty}(x-\mu)^n f(x)dx)$$

Un estimador del momento central de orden n es:

$$m_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Método de de máxima verosimilitud

Considerar x distribuida de acuerdo a f(x;q) donde q es un parámetro (o vector de parámetros) desconocido.

El método de máxima verosimilitud es una técnica para estimar los valores de θ dada una muestra finita de datos.

Supongamos n medidas de x, $x_1,...,x_n$. Puesto que las medidas son independientes, la probabilidad de que x_1 esté en $[x_1,x_1+dx_1]$, x_2 en $[x_2,x_2+dx_2]$, es:

probabilidad de que x_i esté en $[x_i, x_i + dx_i]$para todo $i=\prod_{i=1}^{n} f(x_i;\theta)dx_i$

Si la la pdf y el (los) parámetro(s) describen realmente los datos, esperamos alta probabilidad para los datos que hemos medido. Análogamente un parámetro cuyo valor se desvíe mucho del auténtico resultará en baja probabilidad para las medidas observadas.

Función de verosimilitud

$$P(\text{todo } x_i \text{ en } [x_i, x_i + dx_i]) = \prod_{i=1}^{n} f(x_i;\theta)dx_i$$

Probabilidad máxima para la pdf y parámetros correctos. Por tanto la función:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Será máxima para la pdf y parámetros correctos. En estadística clásica L(θ) no es la pdf de θ sino la pdf conjunta de los x donde:

q se trata como un parámetro (del que la pdf depende)

los xi están fijados (los datos ya han sido adquiridos)

En estadística Bayesiana, podemos tratar L(θ)=L(x|θ) como la pdf de x dado θ y a usar el teorema de Bayes para calcular la probabilidad posterior p(θ|x).

Estimadores de máxima verosimilitud

Se definen los estimadores de máxima verosimilitud de los parámetros como aquellos que maximizan la función de verosimilitud:

$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots m$$

NB: La definición no garantiza que los estimadores MV sean “óptimos” en absoluto! En general, sin embargo, suelen ser la aproximación más aceptable al problema de estimar parámetros.

Ejemplo: Distribución exponencial

Suponer que se han medido los tiempos de desintegración (propios) de una muestra de leptones tau, obteniéndose un conjunto de valores $t_1, t_2, \ldots t_n$.

Escogemos como HIPÓTESIS para la distribución de los t_i una pdf exponencial con media τ.

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

Nuestro objetivo es estimar el valor del parámetro τ. Para ello usamos la función de verosimilitud (de hecho, su logaritmo, más fácil de manejar)

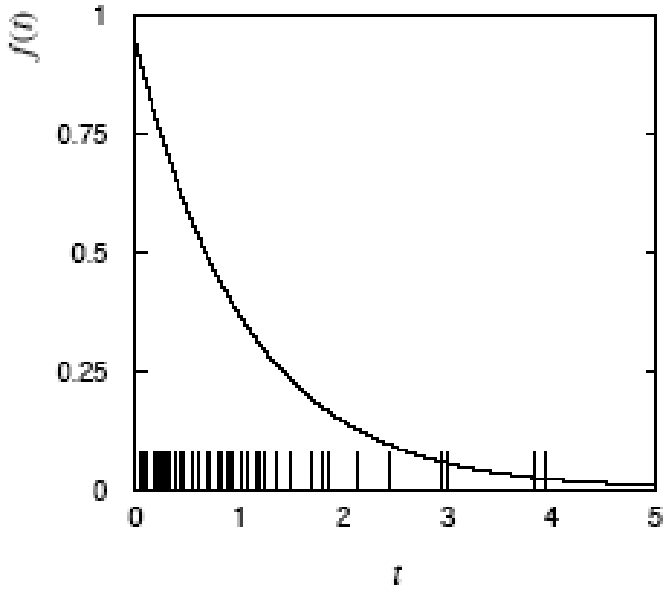
$$\log L(\tau) = \sum_{i=1}^{n} \log f(t_i;\tau) = \sum_{i=1}^{n} (\log \frac{1}{\tau} - \frac{t_i}{\tau})$$

$$\frac{\partial \log L(\tau)}{\partial \tau} = 0 \rightarrow \sum_{i=1}^{n} \tau(-\frac{1}{\tau^2}) + (-t_i)(-\frac{1}{\tau^2}) =$$

$$= -\frac{1}{\tau^2} \sum_{i=1}^{n} \tau - t_i = 0$$

$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Example: generate 50 values of t with MC using $\tau = 1$,



$\hat{\tau} = 1.062$

Valor esperado: $$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i \rightarrow E[\hat{\tau}] = \tau = \frac{1}{n}\sum_{i=1}^{n} \tau$$

inmediato de calcular, puesto que el estimador es la media muestral, cuyo valor esperado coincide con la media poblacional, esto es con τ. Por lo tanto el estimador $\bar{\tau}$ no tiene sesgo.

Supongamos que en lugar de la vida media queremos calcular la constante de desintegración $\lambda = 1/\tau$:

$\lambda = 1/\tau$ sólo es un estimador sin sesgo de τ en el límite de alto n!

Es decir: El estimador MV de una función del parámetro θ, $a=a(\theta)$ no es más que $\bar{a}=a(\bar{\theta})$.

Pero si $\bar{\theta}$ es un estimador sin sesgo de θ no necesariamente a es un estimador sin sesgo de $a(\theta)$

$$\lambda = \lambda(\tau)$$

$$\frac{\partial L}{\partial \lambda} = \frac{\partial L}{\partial \tau}\frac{\partial \tau}{\partial \lambda}$$

$$\frac{\partial L}{\partial \tau} = 0 \rightarrow \frac{\partial L}{\partial \lambda} = 0 \text{ siempre que } \frac{\partial \tau}{\partial \lambda} \neq 0$$

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = \frac{n}{\sum_{i=1}^{n} t_i}$$

$$E[\hat{\lambda}] = \lambda\frac{n}{n-1} = \frac{1}{\tau}\frac{n}{n-1}$$

Ejemplo: Estimadores MV de una gausiana

Considerar n medidas de x que asumimos distribuidas de acuerdo a una pdf gausiana. La función de verosimilitud es:

$$\log L(\mu,\sigma^2) = \sum_{i=1}^{n} \log f(x;\mu,\sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) \right)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}} + \frac{1}{2}\log\frac{1}{\sigma^2} - \frac{-(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \log L}{\partial \mu} = 0 \rightarrow -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x-\mu) = 0 \rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$E[\hat{\mu}] = \mu \rightarrow \hat{\mu}$ no tiene sesgo

$$\frac{\partial \log L}{\partial \sigma^2} = 0 \rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \rightarrow \hat{\sigma}^2$ no tiene sesgo en el límite asintótico

NB la varianza muestral es siempre un estimador sin sesgo

pero no es un estimador MV!

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Varianza de los estimadores MV

La varianza de un estimador $\overline{\theta}$ nos proporciona una medida de la incertidumbre estadística en el conocimiento de dicho estimador. Esto es:

Si repetimos muchas veces el experimento (con n medidas en cada caso) y obtenemos $\overline{\theta}$ cada vez, ¿cuánto se esparcen sus valores --entorno al valor medio, si no hay sesgo--?

Para ello, calculamos la varianza de $\overline{\theta}$.

Técnicas para calcular la varianza :

Analítica (sólo en algunos casos sencillos)

Monte Carlo

Métodos numéricos

Analítica para el caso de la pdf exponencial

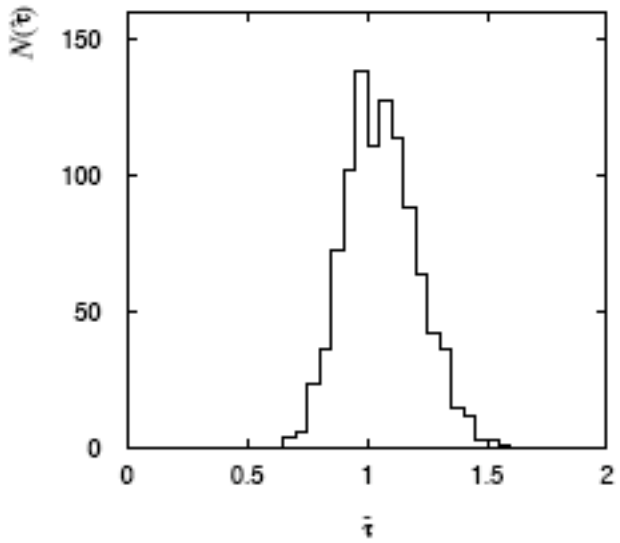
$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2$$

$$= \int \dots \int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \dots dt_n$$

$$- \left(\int \dots \int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \dots dt_n\right)^2$$

$$= \frac{\tau^2}{n}.$$

Monte Carlo: Muchos experimentos, cada uno con n fijo. La varianza viene dada por la dispersión del estimador entorno al valor medio



Desigualdad RCF

Rao-Cramér-Frechet: Para un estimador arbitrario θ se verifica que:

$$\mathrm{V}[\hat{\theta}] \geq \frac{(1+\frac{\partial b}{\partial \theta})^2}{E[-\frac{\partial^2 \log L}{\partial \theta^2}]}$$

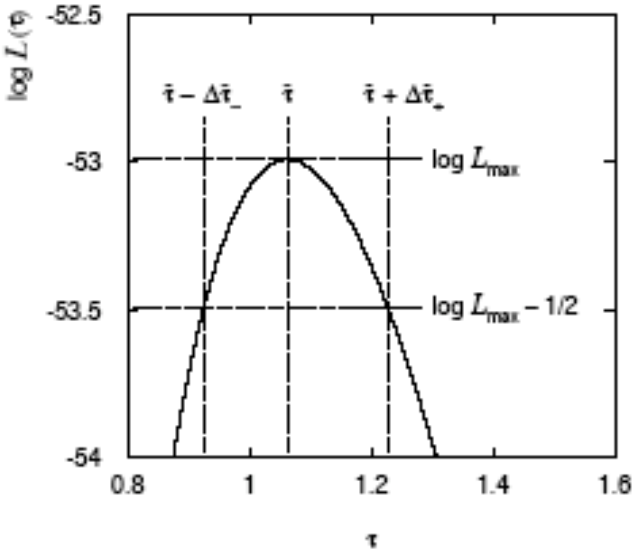
Decimos que un estimador es eficiente (y sin sesgo) cuando se verifica la igualdad estricta. Este es el caso, a menudo con estimadores MV (a veces en el límite asintótico). Otras veces se utiliza la igualdad como una aproximación. En estos casos:

$$\mathrm{V}[\hat{\theta}]^{-1} = E[-\frac{\partial^2 \log L}{\partial \theta^2}], \quad V_{ij}^{-1} = E[-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}]$$

Método numérico para calcular la varianza de un estimador MV

Consideremos el parámetro θ: Expandiendo log (L) en serie de Taylor entorno al estimador $\bar{\theta}$

$$\log L(\theta) = \log L(\hat{\theta}) + \left[\frac{\partial \log L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$



Para obtener σ_θ cambiamos θ alejándonos de q hasta que log L disminuye en 1/2

Alrededor de θ logL es máxima y por tanto la primera derivada se cancela. Utilizando RCF (asumiendo que el estimador es eficiente y sin sesgo)

$$\log L(\theta) = \log L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}^2_{\hat{\theta}}} \rightarrow$$

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{\max} - \frac{1}{2}$$